Personalized cancer diagnosis

1. Business Problem

1.1. Description

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

Data: Memorial Sloan Kettering Cancer Center (MSKCC)

Download training_variants.zip and training_text.zip from Kaggle.

Context:

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

Problem statement :

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

1.2. Source/Useful Links

Some articles and reference blogs about the problem statement

1. https://www.forbes.com/sites/matthewherper/2017/06/03/a-new-cancer-drug-helped-almost-everyone-who-took-it-almost-heres-what-it-teaches-us/#2a44ee2f6b25
2. https://www.youtube.com/watch?v=UwbuW7oK8rk
3. https://www.youtube.com/watch?v=qxXRKVompI8

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 [0]:

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

from mlxtend.classifier import StackingClassifier

from sklearn import model_selection
from sklearn.linear_model import LogisticRegression
import nltk
nltk.download('stopwords')
```

```
/usr/local/lib/python3.6/dist-packages/sklearn/externals/six.py:31: DeprecationWarning: The module is deprecated in version 0.21 and will be removed in version 0.23 since we've dropped support for Python 2.7. Please rely on the official version of six (https://pypi.org/project/six/).
  "(https://pypi.org/project/six/).", DeprecationWarning)
```

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

Out[0]:

```
True
```

Mounting the drive

In [0]:

```
!kill -9 -1
```

In [0]:

```
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=947318989803-6bn6
qk8qdgf4n4g3pfee6491hc0brc4i.apps.googleusercontent.com&redirect_uri=urn%3Aietf%3Awg%3Aoauth%3A2.0%
b&scope=email%20https%3A%2F%2Fwww.googleapis.com%2Fauth%2Fdocs.test%20https%3A%2F%2Fwww.googleapis.
2Fauth%2Fdrive%20https%3A%2F%2Fwww.googleapis.com%2Fauth%2Fdrive.photos.readonly%20https%3A%2F%2Fww
ogleapis.com%2Fauth%2Fpeopleapi.readonly&response_type=code

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

In [0]:

```
!pwd
!ls
```

```
/content
drive  sample_data
```

In [0]:

```
import os
PATH = os.getcwd()
print(PATH)
```

```
/content
```

In [0]:

```
data_path = PATH + '/drive/My Drive/AAIC/Case Studies/Personalized Cancer Diagnosis Case Study/'
data_path
```

Out[0]:

```
'/content/drive/My Drive/AAIC/Case Studies/Personalized Cancer Diagnosis Case Study/'
```

3.1. Reading Data

3.1.1. Reading Gene and Variation Data

In [0]:

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

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

Out[0]:

	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

	ID	Gene	Variation	Class
4	4	CBL	L399V	4

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 [0]:

```
# note the seprator in this file
data_text =pd.read_csv(data_path + "training_text",sep="\|\|",engine="python",names=["ID","TEXT"],s
kiprows=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[0]:

	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 [0]:

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


def nlp_preprocessing(total_text, index, column):
    if type(total_text) is not int:
        string = ""
        # replace every special char with space
        total_text = re.sub('[^a-zA-Z0-9\n]', ' ', total_text)
        # replace multiple spaces with single space
        total_text = re.sub('\s+',' ', total_text)
        # converting all the chars into lower-case.
        total_text = total_text.lower()

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

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

In [0]:

In [0]:

```
#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_time, "seconds")
```

```
there is no text description for id: 1109
there is no text description for id: 1277
there is no text description for id: 1407
there is no text description for id: 1639
there is no text description for id: 2755
Time took for preprocessing the text : 195.62990200000002 seconds
```

In [0]:

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

Out[0]:

	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 [0]:

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

Out[0]:

	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 [0]:

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

Out[0]:

	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 [0]:

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

# split the data into test and train by maintaining same distribution of output varaible 'y_true'
[stratify=y_true]
X_train, test_df, y_train, y_test = train_test_split(result, y_true, stratify=y_true, test_size=0.2
)
# split the train data into train and cross validation by maintaining same distribution of output
varaible 'y_train' [stratify=y_train]
train_df, cv_df, y_train, y_cv = train_test_split(X_train, y_train, stratify=y_train, test_size=0.2
)
```

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

In [0]:

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

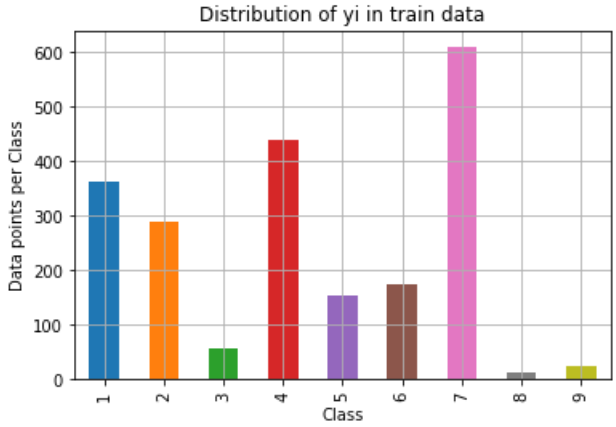
```
Number of data points in train data: 2124
Number of data points in test data: 665
Number of data points in cross validation data: 532
```

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

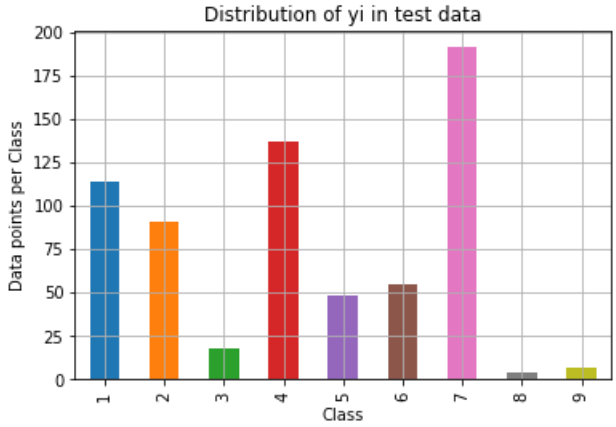
In [0]:

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

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

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


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

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

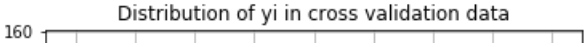
```
print('-'*80)
my_colors = 'rgbkymc'
cv_class_distribution.plot(kind='bar')
plt.xlabel('Class')
plt.ylabel('Data points per Class')
plt.title('Distribution of yi in cross validation data')
plt.grid()
plt.show()

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

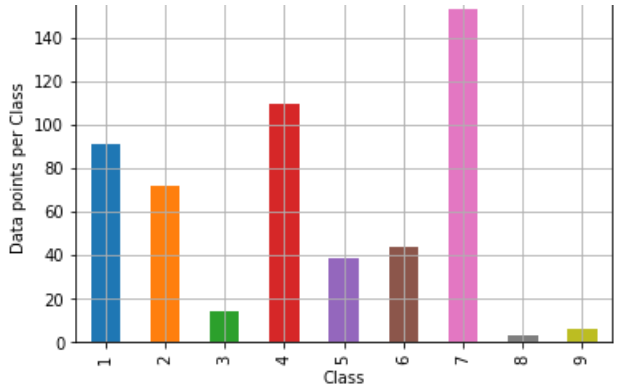



```
Number of data points in class 7 : 609 ( 28.672 %)
Number of data points in class 4 : 439 ( 20.669 %)
Number of data points in class 1 : 363 ( 17.09 %)
Number of data points in class 2 : 289 ( 13.606 %)
Number of data points in class 6 : 176 ( 8.286 %)
Number of data points in class 5 : 155 ( 7.298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




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






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

3.2 Prediction using a 'Random' Model

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

In [0]:

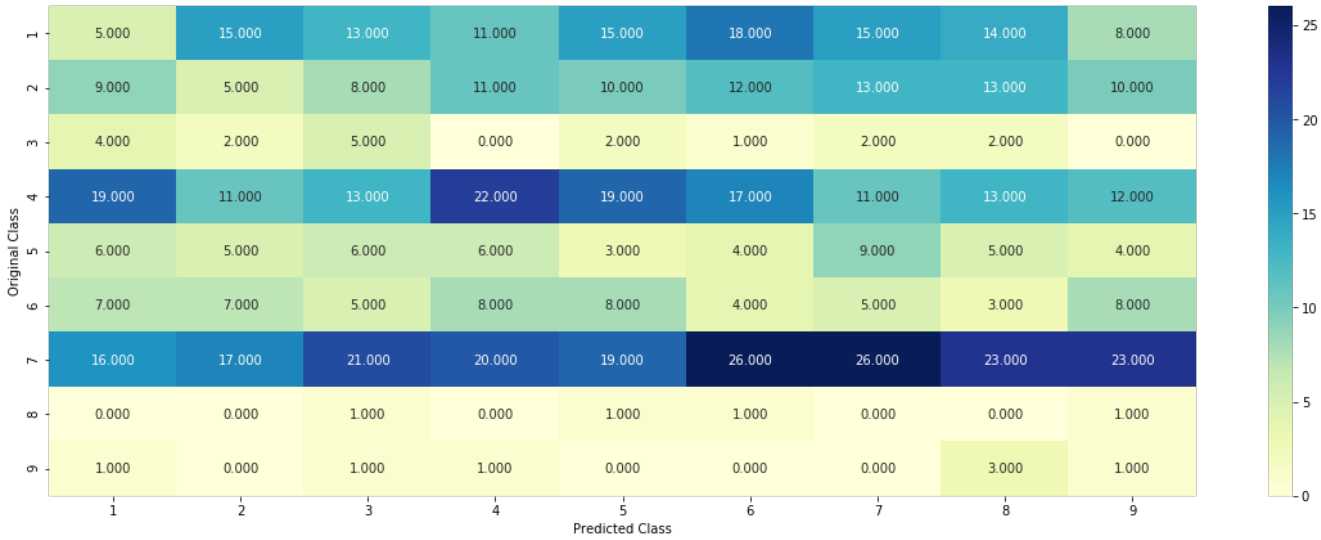
```
# This function plots the confusion matrices given y_i, y_i_hat.
def plot_confusion_matrix(test_y, predict_y):
    C = confusion_matrix(test_y, predict_y)
    # C = 9,9 matrix, each cell (i,j) represents number of points of class i are predicted class j

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

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

    # ((C.T)/(C.sum(axis=1))).T = [[1/3, 2/3]
    #                              [3/7, 4/7]]
    # sum of row elements = 1

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

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

    print("-"*20, "Precision matrix (Columm Sum=1)", "-"*20)
```

```
    plt.figure(figsize=(20,7))
    sns.heatmap(B, annot=True, cmap="YlGnBu", fmt=".3f", xticklabels=labels, yticklabels=labels)
    plt.xlabel('Predicted Class')
    plt.ylabel('Original Class')
    plt.show()

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

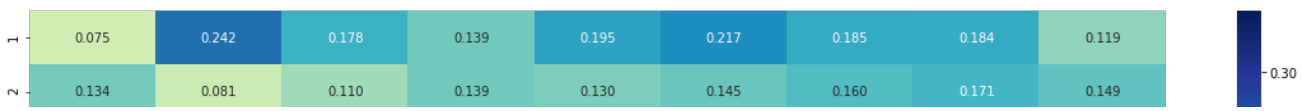
In [0]:

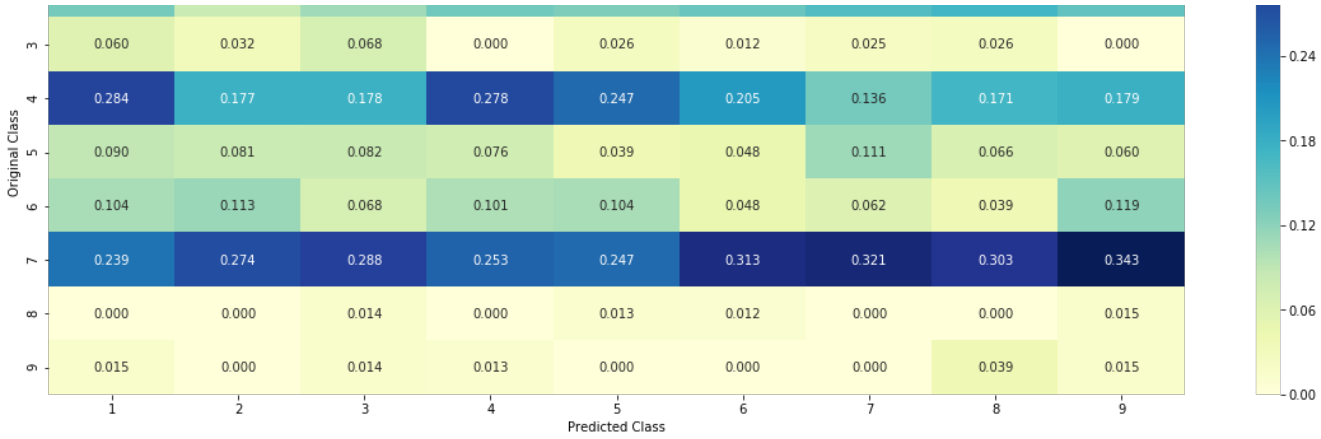
```
# we need to generate 9 numbers and the sum of numbers should be 1
# one solution is to genarate 9 numbers and divide each of the numbers by their sum
# ref: https://stackoverflow.com/a/18662466/4084039
test_data_len = test_df.shape[0]
cv_data_len = cv_df.shape[0]

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


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

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

```
Log loss on Cross Validation Data using Random Model 2.518103352528671
Log loss on Test Data using Random Model 2.4849211935822995
-------------------- 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: ['train_df', 'test_df', 'cv_df']
# algorithm
# ----------
# Consider all unique values and the number of occurances of given feature in train data dataframe
# build a vector (1*9) , the first element = (number of times it occured in class1 + 10*alpha / nu
mber of time it occurred in total data+90*alpha)
# gv_dict is like a look up table, for every gene it store a (1*9) representation of it
# for a value of feature in df:
# if it is in train data:
# we add the vector that was stored in 'gv_dict' look up table to 'gv_fea'
# if it is not there is train:
# we add [1/9, 1/9, 1/9, 1/9,1/9, 1/9, 1/9, 1/9, 1/9] to 'gv_fea'
# return 'gv_fea'
# ----------------------

# get_gv_fea_dict: Get Gene varaition Feature Dict
def get_gv_fea_dict(alpha, feature, df):
    # value_count: it contains a dict like
    # print(train_df['Gene'].value_counts())
    # output:
    #         {BRCA1      174
    #          TP53       106
    #          EGFR        86
    #          BRCA2       75
    #          PTEN        69
    #          KIT         61
```

```
    #         BRAF        60
    #         ERBB2       47
    #         PDGFRA      46
    #         ...}
    # print(train_df['Variation'].value_counts())
    # output:
    # {
    # Truncating_Mutations                     63
    # Deletion                                 43
    # Amplification                            43
    # Fusions                                  22
    # Overexpression                            3
    # E17K                                      3
    # Q61L                                      3
    # S222D                                     2
    # P130S                                     2
    # ...
    # }
    value_count = train_df[feature].value_counts()

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

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

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

            # cls_cnt.shape[0](numerator) will contain the number of time that particular feature o
ccured 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, 0.051020408
163265307, 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, 0.1393939393939394, 0.34545454545454546, 0.060606060606060608,
0.060606060606060608, 0.060606060606060608],
    #      'PTEN': [0.069182389937106917, 0.062893081761006289, 0.069182389937106917,
0.46540880503144655, 0.075471698113207544, 0.062893081761006289, 0.069182389937106917, 0.062893081
761006289, 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 = train_df[feature].value_counts()

    # gv_fea: Gene_variation feature, it will contain the feature for each feature value in the da
ta
    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_fea
    for index, row in df.iterrows():
        if row[feature] in dict(value_count).keys():
            gv_fea.append(gv_dict[row[feature]])
        else:
            gv_fea.append([1/9,1/9,1/9,1/9,1/9,1/9,1/9,1/9,1/9])
#             gv_fea.append([-1,-1,-1,-1,-1,-1,-1,-1,-1])
    return gv_fea
```

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

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

3.2.1 Univariate Analysis on Gene Feature

Q1. Gene, What type of feature it is ?

Ans. Gene is a categorical variable

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

In [0]:

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

```
Number of Unique Genes : 238
BRCA1     171
TP53      107
PTEN       87
EGFR       85
BRCA2      77
KIT        55
BRAF       54
ALK        48
PDGFRA     43
PIK3CA     38
Name: Gene, dtype: int64
```

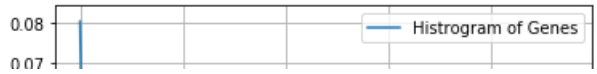
In [0]:

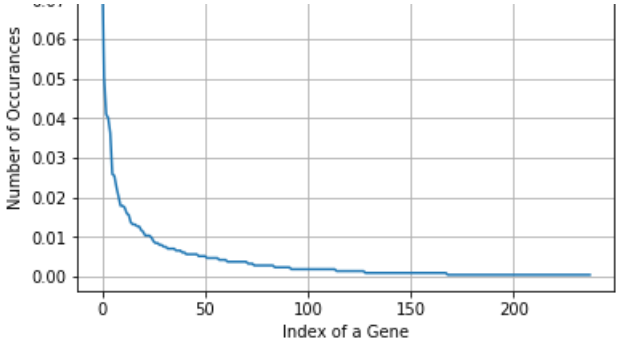
```
print("Ans: There are", unique_genes.shape[0] ,"different categories of genes in the train data, an
d they are distibuted as follows",)
```

```
Ans: There are 238 different categories of genes in the train data, and they are distibuted as fol
lows
```

In [0]:

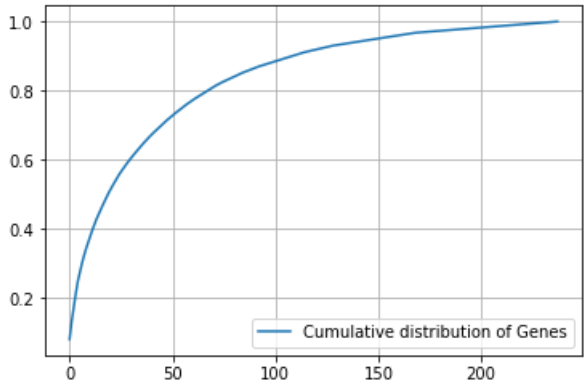
```
s = sum(unique_genes.values);
h = unique_genes.values/s;
plt.plot(h, label="Histrogram of Genes")
plt.xlabel('Index of a Gene')
plt.ylabel('Number of Occurances')
plt.legend()
plt.grid()
plt.show()
```





In [0]:

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

In [0]:

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

```
train_gene_feature_responseCoding is converted feature using respone coding method. The shape of g
ene feature: (2124, 9)
```

In [0]:

In [0]:

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

In [0]:

```
train_df['Gene'].head()
```

Out[0]:

```
1285      HRAS
1388     FGFR1
904     PDGFRA
748      ERBB2
642     CDKN2A
Name: Gene, dtype: object
```

In [0]:

```
gene_vectorizer.get_feature_names()
```

Out[0]:

```
['abl1',
 'acvr1',
 'ago2',
 'akt1',
 'akt2',
 'akt3',
 'alk',
 'apc',
 'ar',
 'araf',
 'arid1b',
 'arid2',
 'arid5b',
 'atm',
 'atrx',
 'aurka',
 'axin1',
 'axl',
 'bap1',
 'bard1',
 'bcl10',
 'bcl2',
 'bcl2l11',
 'bcor',
 'braf',
 'brca1',
 'brca2',
 'brip1',
 'card11',
 'carm1',
 'casp8',
 'cbl',
 'ccnd1',
 'ccnd2',
 'ccnd3',
 'ccne1',
 'cdh1',
 'cdk12',
 'cdk4',
 'cdk6',
 'cdk8',
 'cdkn1a',
 'cdkn2a',
 'cdkn2b',
 'cdkn2c',
 'cebpa',
 'chek2',
 'cic',
```

```
'crebbp',
'ctcf',
'ctla4',
'ctnnb1',
'ddr2',
'dicer1',
'dnmt3a',
'dnmt3b',
'dusp4',
'egfr',
'eif1ax',
'elf3',
'ep300',
'epas1',
'epcam',
'erbb2',
'erbb3',
'erbb4',
'ercc2',
'ercc3',
'ercc4',
'erg',
'errfi1',
'esr1',
'etv1',
'etv6',
'ewsr1',
'ezh2',
'fam58a',
'fanca',
'fancc',
'fat1',
'fbxw7',
'fgf19',
'fgf4',
'fgfr1',
'fgfr2',
'fgfr3',
'fgfr4',
'flt1',
'flt3',
'foxa1',
'foxl2',
'foxo1',
'foxp1',
'fubp1',
'gata3',
'gna11',
'gnaq',
'gnas',
'h3f3a',
'hist1h1c',
'hla',
'hnf1a',
'hras',
'idh1',
'idh2',
'igf1r',
'ikbke',
'jak1',
'jak2',
'jun',
'kdm5c',
'kdm6a',
'kdr',
'keap1',
'kit',
'klf4',
'kmt2a',
'kmt2b',
'kmt2c',
'kmt2d',
'knstrn',
'kras',
'lats2',
'map2k1',
'map2k2',
```

```
'map2k4',
'map3k1',
'mapk1',
'mdm2',
'mdm4',
'med12',
'mef2b',
'men1',
'met',
'mga',
'mlh1',
'mpl',
'msh2',
'msh6',
'mtor',
'myc',
'mycn',
'myd88',
'myod1',
'ncor1',
'nf1',
'nf2',
'nfe2l2',
'nfkbia',
'nkx2',
'notch1',
'notch2',
'nras',
'nsd1',
'ntrk1',
'ntrk2',
'ntrk3',
'nup93',
'pak1',
'pax8',
'pbrm1',
'pdgfra',
'pdgfrb',
'pik3ca',
'pik3cb',
'pik3cd',
'pik3r1',
'pik3r2',
'pik3r3',
'pim1',
'pms1',
'pms2',
'pole',
'ppm1d',
'ppp2r1a',
'ppp6c',
'prdm1',
'ptch1',
'pten',
'ptpn11',
'ptprd',
'ptprt',
'rab35',
'rac1',
'rad21',
'rad50',
'rad51b',
'rad51c',
'rad54l',
'raf1',
'rara',
'rasa1',
'rb1',
'rbm10',
'ret',
'rheb',
'rhoa',
'rit1',
'ros1',
'runx1',
'rxra',
'sdhb',
```

```
 'setd2',
 'sf3b1',
 'shq1',
 'smad2',
 'smad3',
 'smad4',
 'smarca4',
 'smarcb1',
 'smo',
 'sos1',
 'sox9',
 'spop',
 'src',
 'srsf2',
 'stag2',
 'stat3',
 'stk11',
 'tcf3',
 'tcf7l2',
 'tert',
 'tet2',
 'tgfbr1',
 'tgfbr2',
 'tmprss2',
 'tp53',
 'tp53bp1',
 'tsc1',
 'tsc2',
 'u2af1',
 'vegfa',
 'vhl',
 'whsc1l1',
 'xpo1',
 'xrcc2',
 'yap1']
```

In [0]:

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

```
train_gene_feature_onehotCoding is converted feature using one-hot encoding method. The shape of g
ene feature: (2124, 237)
```

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

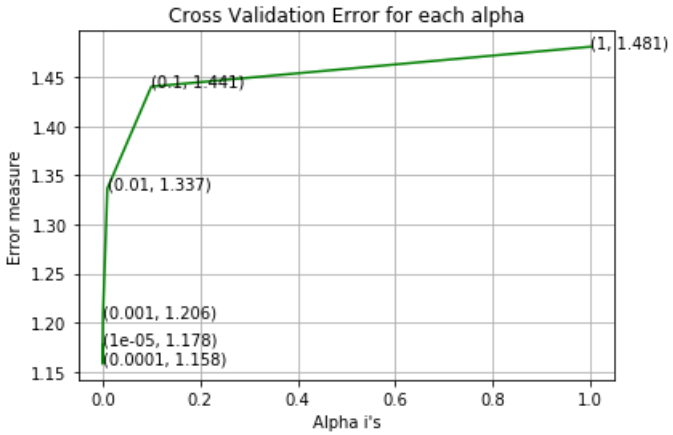
There are many ways to estimate how good a feature is, in predicting y_i. One of the good methods is to build a proper ML model using just this feature. In this case, we will build a logistic regression model using only Gene feature (one hot encoded) to predict y_i.

In [0]:

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

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

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

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


cv_log_error_array=[]
```

```
for i in alpha:
    clf = SGDClassifier(alpha=i, penalty='l2', loss='log', random_state=42)
    clf.fit(train_gene_feature_onehotCoding, y_train)
    sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
    sig_clf.fit(train_gene_feature_onehotCoding, y_train)
    predict_y = sig_clf.predict_proba(cv_gene_feature_onehotCoding)
    cv_log_error_array.append(log_loss(y_cv, predict_y, labels=clf.classes_, eps=1e-15))
    print('For values of alpha = ', i, "The log loss is:",log_loss(y_cv, predict_y, labels=clf.clas
ses_, eps=1e-15))

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


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

predict_y = sig_clf.predict_proba(train_gene_feature_onehotCoding)
print('For values of best alpha = ', alpha[best_alpha], "The train log loss is:",log_loss(y_train,
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_lo
ss(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, p
redict_y, labels=clf.classes_, eps=1e-15))
```

```
For values of alpha =  1e-05 The log loss is: 1.1775994152897509
For values of alpha =  0.0001 The log loss is: 1.1578830894239573
For values of alpha =  0.001 The log loss is: 1.2056464444001167
For values of alpha =  0.01 The log loss is: 1.3366596669556732
For values of alpha =  0.1 The log loss is: 1.4405372030153103
For values of alpha =  1 The log loss is: 1.4811514790413116
```




```
For values of best alpha =  0.0001 The train log loss is: 1.0013546499232338
For values of best alpha =  0.0001 The cross validation log loss is: 1.1578830894239573
For values of best alpha =  0.0001 The test log loss is: 1.2085435525874688
```

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 [0]:

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

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

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

```
Q6. How many data points in Test and CV datasets are covered by the  238  genes in train dataset?
Ans
1. In test data 644 out of 665 : 96.84210526315789
2. In cross validation data 513 out of  532 : 96.42857142857143
```

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 [0]:

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

```
Number of Unique Variations : 1936
Truncating_Mutations    55
Amplification           47
Deletion                44
Fusions                 22
Overexpression           3
Q209L                    2
G12C                     2
G12V                     2
Y64A                     2
Q61H                     2
Name: Variation, dtype: int64
```

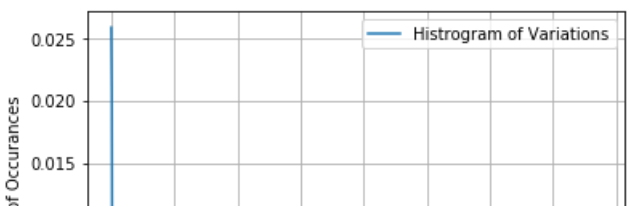
In [0]:

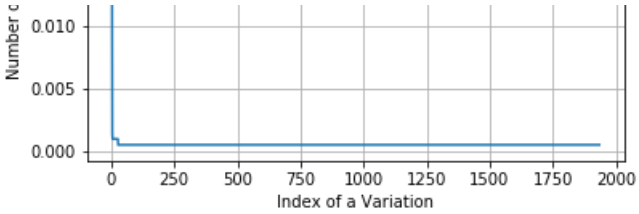
```
print("Ans: There are", unique_variations.shape[0] ,"different categories of variations in the
train data, and they are distibuted as follows",)
```

```
Ans: There are 1936 different categories of variations in the train data, and they are distibuted
as follows
```

In [0]:

```
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()
```

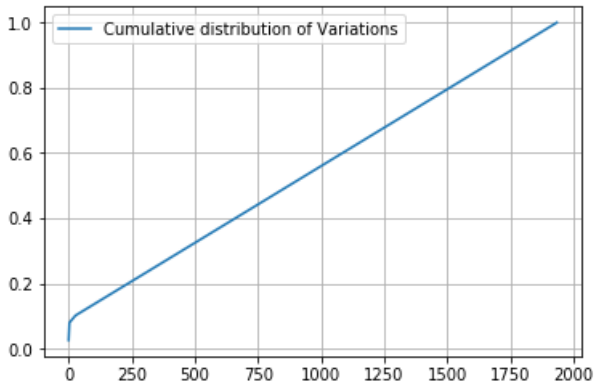




In [0]:

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

```
[0.02589454 0.0480226  0.06873823 ... 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", train_df))
# test gene feature
test_variation_feature_responseCoding = np.array(get_gv_feature(alpha, "Variation", test_df))
# cross validation gene feature
cv_variation_feature_responseCoding = np.array(get_gv_feature(alpha, "Variation", cv_df))
```

In [0]:

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

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

In [0]:

```
# one-hot encoding of variation feature.
variation_vectorizer = CountVectorizer()
```

```
train_variation_feature_onehotCoding = variation_vectorizer.fit_transform(train_df['Variation'])
test_variation_feature_onehotCoding = variation_vectorizer.transform(test_df['Variation'])
cv_variation_feature_onehotCoding = variation_vectorizer.transform(cv_df['Variation'])
```

In [0]:

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

```
train_variation_feature_onehotEncoded is converted feature using the onne-hot encoding method. The
shape of Variation feature: (2124, 1971)
```

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

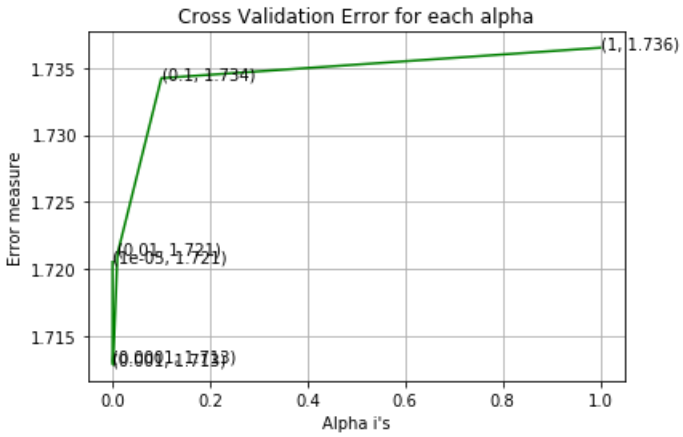
Let's build a model just like the earlier!

In [0]:

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

# read more about SGDClassifier() at http://scikit-
learn.org/stable/modules/generated/sklearn.linear_model.SGDClassifier.html
# ------------------------------
# default parameters
# SGDClassifier(loss='hinge', penalty='l2', alpha=0.0001, l1_ratio=0.15, fit_intercept=True, max_i
ter=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.clas
ses_, 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 validation log loss is:",log_lo
ss(y_cv, predict_y, labels=clf.classes_, eps=1e-15))
predict_y = sig_clf.predict_proba(test_variation_feature_onehotCoding)
print('For values of best alpha = ', alpha[best_alpha], "The test log loss is:",log_loss(y_test, p
redict_y, labels=clf.classes_, eps=1e-15))
```

```
For values of alpha =  1e-05 The log loss is: 1.7205334432745771
For values of alpha =  0.0001 The log loss is: 1.7131754324463908
For values of alpha =  0.001 The log loss is: 1.7128725944334389
For values of alpha =  0.01 The log loss is: 1.7211382388742982
For values of alpha =  0.1 The log loss is: 1.7342235532451458
For values of alpha =  1 The log loss is: 1.7364863281773133
```




```
For values of best alpha =  0.001 The train log loss is: 1.0757847732262715
For values of best alpha =  0.001 The cross validation log loss is: 1.7128725944334389
For values of best alpha =  0.001 The test log loss is: 1.6973775544342997
```

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 [0]:

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

```
Q12. How many data points are covered by total  1936  genes in test and cross validation data
sets?
Ans
1. In test data 68 out of 665 : 10.225563909774436
2. In cross validation data 63 out of  532 : 11.842105263157894
```

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 )/(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 [0]:

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

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

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


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

```
Total number of unique words in train data : 52114
```

In [0]:

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

# dict_list[i] is build on i'th  class text data
# total_dict is buid on whole training text data
total_dict = extract_dictionary_paddle(train_df)


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(train_df)
test_text_feature_responseCoding  = get_text_responsecoding(test_df)
cv_text_feature_responseCoding  = get_text_responsecoding(cv_df)
```

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(test_df['TEXT'])
# don't forget to normalize every feature
test_text_feature_onehotCoding = normalize(test_text_feature_onehotCoding, axis=0)

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

In [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 [0]:

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

```
Counter({3: 4550, 4: 3678, 5: 2783, 6: 2466, 8: 2195, 7: 2186, 11: 1400, 10: 1336, 9: 1316, 12: 129
1, 15: 1030, 14: 1020, 13: 914, 16: 779, 20: 614, 18: 589, 24: 582, 19: 567, 17: 552, 21: 540, 22:
464, 40: 391, 25: 390, 23: 371, 26: 366, 30: 353, 27: 353, 28: 336, 32: 289, 29: 284, 50: 282, 33:
279, 36: 273, 34: 272, 31: 268, 35: 257, 42: 249, 48: 238, 37: 217, 44: 198, 39: 194, 38: 183, 41:
182, 45: 175, 43: 172, 49: 163, 55: 162, 58: 159, 56: 153, 54: 151, 46: 148, 51: 147, 60: 142, 47:
139, 57: 136, 53: 133, 52: 129, 62: 120, 66: 113, 59: 113, 65: 101, 61: 100, 69: 99, 63: 99, 64:
98, 72: 96, 84: 93, 74: 92, 78: 89, 96: 88, 71: 88, 70: 86, 67: 84, 80: 82, 79: 81, 75: 81, 68: 7
7, 83: 75, 93: 74, 77: 73, 73: 72, 91: 70, 86: 69, 105: 68, 82: 68, 81: 68, 100: 67, 76: 67, 88:
65, 98: 64, 87: 64, 95: 63, 120: 62, 112: 62, 103: 62, 90: 62, 92: 60, 89: 58, 101: 56, 85: 56, 1
10: 55, 102: 52, 99: 52, 94: 52, 108: 51, 104: 51, 150: 49, 119: 49, 117: 49, 114: 49, 97: 48,
142: 45, 109: 44, 130: 43, 144: 42, 126: 42, 136: 41, 123: 41, 156: 40, 138: 40, 135: 40, 129: 40,
147: 39, 141: 39, 107: 39, 132: 38, 127: 38, 151: 37, 115: 37, 111: 37, 133: 36, 168: 35, 143: 35,
140: 35, 137: 35, 113: 35, 161: 34, 134: 34, 125: 34, 106: 34, 124: 33, 152: 32, 148: 32, 139: 32,
128: 32, 121: 32, 116: 31, 180: 30, 131: 30, 183: 29, 177: 29, 153: 29, 146: 29, 170: 28, 118: 28,
204: 27, 232: 26, 220: 26, 203: 26, 196: 26, 181: 26, 166: 26, 162: 26, 159: 26, 149: 26, 230: 25,
216: 25, 178: 25, 201: 24, 199: 24, 184: 24, 155: 24, 154: 24, 297: 23, 231: 23, 194: 23, 175: 23,
212: 22, 205: 22, 195: 22, 169: 22, 165: 22, 158: 22, 157: 22, 202: 21, 189: 21, 188: 21, 176: 21,
172: 21, 160: 21, 145: 21, 122: 21, 295: 20, 255: 20, 218: 20, 211: 20, 210: 20, 200: 20, 185: 20,
174: 20, 173: 20, 317: 19, 213: 19, 197: 19, 179: 19, 164: 19, 163: 19, 367: 18, 300: 18, 238: 18,
235: 18, 208: 18, 193: 18, 192: 18, 187: 18, 171: 18, 351: 17, 334: 17, 288: 17, 252: 17, 243: 17,
240: 17, 239: 17, 206: 17, 198: 17, 190: 17, 186: 17, 182: 17, 328: 16, 308: 16, 276: 16, 251: 16,
247: 16, 233: 16, 223: 16, 219: 16, 290: 15, 278: 15, 273: 15, 258: 15, 254: 15, 241: 15, 237: 15,
234: 15, 224: 15, 221: 15, 207: 15, 191: 15, 376: 14, 319: 14, 305: 14, 294: 14, 274: 14, 263: 14,
262: 14, 261: 14, 259: 14, 257: 14, 245: 14, 226: 14, 215: 14, 167: 14, 383: 13, 355: 13, 341: 13,
324: 13, 306: 13, 302: 13, 289: 13, 285: 13, 281: 13, 280: 13, 277: 13, 268: 13, 256: 13, 250: 13,
249: 13, 228: 13, 222: 13, 214: 13, 352: 12, 321: 12, 312: 12, 296: 12, 292: 12, 287: 12, 272: 12,
264: 12, 260: 12, 253: 12, 246: 12, 244: 12, 236: 12, 227: 12, 225: 12, 209: 12, 407: 11, 373: 11,
371: 11, 353: 11, 349: 11, 343: 11, 337: 11, 329: 11, 323: 11, 322: 11, 270: 11, 655: 10, 490: 10,
401: 10, 387: 10, 375: 10, 356: 10, 340: 10, 338: 10, 336: 10, 331: 10, 309: 10, 301: 10, 283: 10,
282: 10, 279: 10, 269: 10, 267: 10, 242: 10, 217: 10, 788: 9, 466: 9, 441: 9, 424: 9, 404: 9, 400:
9, 391: 9, 365: 9, 362: 9, 354: 9, 318: 9, 313: 9, 311: 9, 299: 9, 298: 9, 284: 9, 271: 9, 266: 9
```

9, 391: 9, 385: 9, 382: 9, 334: 9, 316: 9, 315: 9, 311: 9, 299: 9, 290: 9, 284: 9, 271: 9, 260: 9,
229: 9, 652: 8, 604: 8, 582: 8, 547: 8, 530: 8, 517: 8, 500: 8, 498: 8, 468: 8, 456: 8, 437: 8,
434: 8, 430: 8, 426: 8, 425: 8, 409: 8, 398: 8, 397: 8, 394: 8, 393: 8, 364: 8, 350: 8, 326: 8,
314: 8, 275: 8, 265: 8, 248: 8, 645: 7, 639: 7, 631: 7, 611: 7, 585: 7, 579: 7, 574: 7, 555: 7,
503: 7, 494: 7, 480: 7, 473: 7, 464: 7, 462: 7, 461: 7, 460: 7, 452: 7, 450: 7, 449: 7, 448: 7,
428: 7, 422: 7, 417: 7, 408: 7, 392: 7, 388: 7, 382: 7, 381: 7, 380: 7, 378: 7, 360: 7, 358: 7,
339: 7, 332: 7, 330: 7, 320: 7, 315: 7, 310: 7, 307: 7, 1303: 6, 851: 6, 808: 6, 726: 6, 722: 6,
696: 6, 692: 6, 686: 6, 666: 6, 661: 6, 638: 6, 615: 6, 602: 6, 572: 6, 561: 6, 559: 6, 557: 6,
553: 6, 544: 6, 531: 6, 529: 6, 510: 6, 508: 6, 507: 6, 501: 6, 483: 6, 477: 6, 475: 6, 470: 6,
463: 6, 458: 6, 453: 6, 446: 6, 440: 6, 435: 6, 427: 6, 418: 6, 415: 6, 414: 6, 411: 6, 406: 6,
396: 6, 390: 6, 377: 6, 368: 6, 348: 6, 347: 6, 344: 6, 342: 6, 325: 6, 316: 6, 303: 6, 293: 6,
291: 6, 286: 6, 1257: 5, 989: 5, 943: 5, 939: 5, 934: 5, 931: 5, 924: 5, 846: 5, 825: 5, 792: 5,
767: 5, 755: 5, 744: 5, 737: 5, 733: 5, 711: 5, 702: 5, 690: 5, 667: 5, 650: 5, 648: 5, 642: 5,
640: 5, 635: 5, 632: 5, 628: 5, 627: 5, 610: 5, 608: 5, 594: 5, 588: 5, 575: 5, 565: 5, 550: 5,
542: 5, 539: 5, 527: 5, 524: 5, 523: 5, 516: 5, 513: 5, 506: 5, 505: 5, 504: 5, 499: 5, 497: 5,
495: 5, 492: 5, 488: 5, 481: 5, 476: 5, 471: 5, 457: 5, 454: 5, 445: 5, 436: 5, 433: 5, 423: 5,
420: 5, 410: 5, 405: 5, 403: 5, 389: 5, 386: 5, 384: 5, 379: 5, 374: 5, 372: 5, 366: 5, 361: 5,
345: 5, 333: 5, 304: 5, 2432: 4, 2004: 4, 2002: 4, 1853: 4, 1562: 4, 1553: 4, 1536: 4, 1534: 4, 145
2: 4, 1430: 4, 1404: 4, 1376: 4, 1360: 4, 1335: 4, 1333: 4, 1291: 4, 1272: 4, 1266: 4, 1239: 4, 120
0: 4, 1199: 4, 1172: 4, 1120: 4, 1051: 4, 1037: 4, 1010: 4, 1003: 4, 995: 4, 993: 4, 979: 4, 966: 4
, 941: 4, 933: 4, 927: 4, 925: 4, 895: 4, 883: 4, 873: 4, 872: 4, 866: 4, 865: 4, 856: 4, 847: 4,
841: 4, 836: 4, 826: 4, 820: 4, 814: 4, 809: 4, 804: 4, 784: 4, 776: 4, 771: 4, 762: 4, 761: 4,
741: 4, 732: 4, 709: 4, 698: 4, 685: 4, 678: 4, 669: 4, 663: 4, 662: 4, 629: 4, 625: 4, 623: 4,
621: 4, 620: 4, 618: 4, 617: 4, 616: 4, 605: 4, 603: 4, 600: 4, 596: 4, 592: 4, 589: 4, 581: 4,
577: 4, 573: 4, 567: 4, 563: 4, 560: 4, 554: 4, 549: 4, 546: 4, 543: 4, 541: 4, 538: 4, 518: 4,
512: 4, 511: 4, 496: 4, 484: 4, 482: 4, 469: 4, 455: 4, 451: 4, 447: 4, 444: 4, 442: 4, 432: 4,
431: 4, 429: 4, 416: 4, 413: 4, 412: 4, 399: 4, 385: 4, 370: 4, 369: 4, 346: 4, 327: 4, 4640: 3,
2659: 3, 2651: 3, 2550: 3, 2479: 3, 2427: 3, 2400: 3, 2313: 3, 2214: 3, 2084: 3, 1998: 3, 1987: 3,
1984: 3, 1928: 3, 1890: 3, 1861: 3, 1831: 3, 1785: 3, 1780: 3, 1779: 3, 1757: 3, 1660: 3, 1649: 3,
1644: 3, 1611: 3, 1587: 3, 1568: 3, 1552: 3, 1550: 3, 1544: 3, 1531: 3, 1499: 3, 1478: 3, 1437: 3,
1422: 3, 1409: 3, 1408: 3, 1361: 3, 1357: 3, 1345: 3, 1337: 3, 1317: 3, 1307: 3, 1288: 3, 1287: 3,
1279: 3, 1265: 3, 1251: 3, 1238: 3, 1222: 3, 1215: 3, 1196: 3, 1187: 3, 1186: 3, 1179: 3, 1153: 3,
1148: 3, 1134: 3, 1129: 3, 1126: 3, 1118: 3, 1116: 3, 1077: 3, 1064: 3, 1061: 3, 1047: 3, 1045: 3,
1042: 3, 1038: 3, 1034: 3, 1017: 3, 1016: 3, 983: 3, 980: 3, 977: 3, 971: 3, 970: 3, 959: 3, 957: 3
, 954: 3, 948: 3, 946: 3, 938: 3, 926: 3, 922: 3, 919: 3, 909: 3, 902: 3, 894: 3, 886: 3, 879: 3,
870: 3, 868: 3, 864: 3, 861: 3, 849: 3, 844: 3, 842: 3, 840: 3, 835: 3, 831: 3, 829: 3, 827: 3,
819: 3, 815: 3, 813: 3, 807: 3, 798: 3, 797: 3, 795: 3, 791: 3, 790: 3, 789: 3, 786: 3, 781: 3,
777: 3, 775: 3, 774: 3, 769: 3, 768: 3, 763: 3, 758: 3, 756: 3, 754: 3, 753: 3, 752: 3, 749: 3,
738: 3, 735: 3, 729: 3, 728: 3, 725: 3, 723: 3, 717: 3, 714: 3, 710: 3, 701: 3, 699: 3, 697: 3,
694: 3, 688: 3, 687: 3, 683: 3, 682: 3, 680: 3, 679: 3, 676: 3, 675: 3, 674: 3, 670: 3, 665: 3,
664: 3, 651: 3, 647: 3, 646: 3, 644: 3, 637: 3, 630: 3, 612: 3, 609: 3, 601: 3, 597: 3, 587: 3,
584: 3, 583: 3, 580: 3, 578: 3, 576: 3, 570: 3, 569: 3, 568: 3, 558: 3, 556: 3, 552: 3, 551: 3,
548: 3, 540: 3, 537: 3, 535: 3, 533: 3, 532: 3, 528: 3, 521: 3, 519: 3, 515: 3, 514: 3, 509: 3,
502: 3, 485: 3, 474: 3, 472: 3, 467: 3, 459: 3, 443: 3, 439: 3, 438: 3, 421: 3, 419: 3, 402: 3,
395: 3, 359: 3, 335: 3, 11978: 2, 6664: 2, 6205: 2, 6032: 2, 5818: 2, 5672: 2, 5521: 2, 5385: 2, 50
24: 2, 4683: 2, 4347: 2, 4258: 2, 4221: 2, 4198: 2, 4118: 2, 4077: 2, 4020: 2, 3966: 2, 3959: 2, 37
92: 2, 3767: 2, 3753: 2, 3729: 2, 3652: 2, 3635: 2, 3634: 2, 3568: 2, 3564: 2, 3557: 2, 3530: 2, 35
10: 2, 3487: 2, 3427: 2, 3392: 2, 3372: 2, 3367: 2, 3328: 2, 3298: 2, 3297: 2, 3270: 2, 3233: 2, 31
97: 2, 3186: 2, 3158: 2, 3084: 2, 3046: 2, 3040: 2, 3018: 2, 2991: 2, 2930: 2, 2919: 2, 2759: 2, 26
87: 2, 2641: 2, 2632: 2, 2629: 2, 2605: 2, 2595: 2, 2594: 2, 2590: 2, 2569: 2, 2562: 2, 2546: 2, 25
45: 2, 2535: 2, 2507: 2, 2498: 2, 2490: 2, 2468: 2, 2438: 2, 2406: 2, 2402: 2, 2396: 2, 2350: 2, 23
48: 2, 2297: 2, 2282: 2, 2270: 2, 2194: 2, 2166: 2, 2165: 2, 2163: 2, 2161: 2, 2152: 2, 2128: 2, 21
20: 2, 2101: 2, 2099: 2, 2092: 2, 2086: 2, 2065: 2, 2054: 2, 2051: 2, 2020: 2, 2014: 2, 2012: 2, 19
93: 2, 1990: 2, 1981: 2, 1976: 2, 1974: 2, 1970: 2, 1956: 2, 1955: 2, 1951: 2, 1945: 2, 1918: 2, 19
17: 2, 1912: 2, 1909: 2, 1894: 2, 1892: 2, 1877: 2, 1869: 2, 1850: 2, 1844: 2, 1836: 2, 1835: 2, 18
34: 2, 1828: 2, 1826: 2, 1825: 2, 1800: 2, 1798: 2, 1795: 2, 1782: 2, 1764: 2, 1752: 2, 1719: 2, 17
15: 2, 1699: 2, 1694: 2, 1690: 2, 1687: 2, 1685: 2, 1677: 2, 1656: 2, 1652: 2, 1623: 2, 1615: 2, 16
09: 2, 1591: 2, 1572: 2, 1556: 2, 1551: 2, 1548: 2, 1543: 2, 1542: 2, 1541: 2, 1540: 2, 1539: 2, 15
30: 2, 1523: 2, 1520: 2, 1518: 2, 1509: 2, 1502: 2, 1487: 2, 1482: 2, 1471: 2, 1469: 2, 1466: 2, 14
62: 2, 1444: 2, 1415: 2, 1411: 2, 1407: 2, 1399: 2, 1398: 2, 1395: 2, 1393: 2, 1379: 2, 1378: 2, 13
74: 2, 1371: 2, 1367: 2, 1366: 2, 1356: 2, 1336: 2, 1329: 2, 1323: 2, 1322: 2, 1310: 2, 1304: 2, 12
98: 2, 1293: 2, 1290: 2, 1285: 2, 1282: 2, 1281: 2, 1277: 2, 1274: 2, 1269: 2, 1264: 2, 1263: 2, 12
61: 2, 1256: 2, 1255: 2, 1250: 2, 1249: 2, 1248: 2, 1241: 2, 1240: 2, 1232: 2, 1227: 2, 1224: 2, 12
23: 2, 1220: 2, 1216: 2, 1210: 2, 1206: 2, 1205: 2, 1204: 2, 1203: 2, 1202: 2, 1195: 2, 1194: 2, 11
93: 2, 1192: 2, 1181: 2, 1180: 2, 1177: 2, 1176: 2, 1174: 2, 1168: 2, 1165: 2, 1160: 2, 1156: 2, 11
54: 2, 1152: 2, 1149: 2, 1147: 2, 1142: 2, 1141: 2, 1136: 2, 1133: 2, 1127: 2, 1123: 2, 1112: 2, 11
06: 2, 1104: 2, 1103: 2, 1100: 2, 1098: 2, 1096: 2, 1093: 2, 1092: 2, 1091: 2, 1082: 2, 1059: 2, 10
55: 2, 1048: 2, 1040: 2, 1036: 2, 1035: 2, 1032: 2, 1023: 2, 1021: 2, 1020: 2, 1018: 2, 1015: 2, 10
12: 2, 1009: 2, 1008: 2, 1007: 2, 1004: 2, 985: 2, 984: 2, 982: 2, 975: 2, 974: 2, 972: 2, 961: 2,
958: 2, 953: 2, 949: 2, 936: 2, 932: 2, 929: 2, 928: 2, 921: 2, 915: 2, 910: 2, 907: 2, 905: 2,
904: 2, 903: 2, 901: 2, 896: 2, 885: 2, 884: 2, 880: 2, 876: 2, 875: 2, 871: 2, 869: 2, 867: 2,
863: 2, 860: 2, 857: 2, 845: 2, 839: 2, 838: 2, 837: 2, 834: 2, 833: 2, 828: 2, 824: 2, 823: 2,
822: 2, 817: 2, 812: 2, 811: 2, 810: 2, 806: 2, 801: 2, 799: 2, 787: 2, 783: 2, 778: 2, 773: 2,
772: 2, 770: 2, 759: 2, 748: 2, 747: 2, 746: 2, 745: 2, 742: 2, 739: 2, 736: 2, 731: 2, 730: 2,
721: 2, 719: 2, 716: 2, 713: 2, 708: 2, 704: 2, 695: 2, 689: 2, 684: 2, 677: 2, 673: 2, 671: 2,
668: 2, 658: 2, 653: 2, 649: 2, 643: 2, 641: 2, 634: 2, 633: 2, 619: 2, 614: 2, 613: 2, 607: 2,
606: 2, 599: 2, 595: 2, 586: 2, 566: 2, 564: 2, 562: 2, 536: 2, 534: 2, 525: 2, 522: 2, 493: 2,
491: 2, 486: 2, 479: 2, 465: 2, 363: 2, 152413: 1, 117791: 1, 80455: 1, 68189: 1, 68160: 1, 67280:

491: 2, 486: 2, 479: 2, 465: 2, 365: 2, 152415: 1, 117791: 1, 80455: 1, 68169: 1, 68160: 1, 67260:
1, 66849: 1, 62970: 1, 62230: 1, 54692: 1, 53130: 1, 49684: 1, 49167: 1, 46912: 1, 45858: 1, 44458
: 1, 42376: 1, 42216: 1, 41375: 1, 40929: 1, 40706: 1, 39790: 1, 39720: 1, 38301: 1, 37869: 1, 376
91: 1, 36163: 1, 35608: 1, 35407: 1, 34199: 1, 34140: 1, 33528: 1, 33200: 1, 33144: 1, 32494: 1, 3
1542: 1, 29383: 1, 29029: 1, 27848: 1, 26587: 1, 26115: 1, 26039: 1, 25512: 1, 25298: 1, 25165: 1,
25035: 1, 24554: 1, 24538: 1, 24527: 1, 24207: 1, 23944: 1, 23416: 1, 22697: 1, 22629: 1, 22444: 1
, 22360: 1, 21920: 1, 21852: 1, 21811: 1, 21227: 1, 20857: 1, 20614: 1, 20593: 1, 20355: 1, 20086:
1, 20068: 1, 19975: 1, 19756: 1, 19426: 1, 19411: 1, 19278: 1, 18974: 1, 18969: 1, 18789: 1, 18769
: 1, 18728: 1, 18419: 1, 18335: 1, 18202: 1, 18043: 1, 17980: 1, 17867: 1, 17818: 1, 17766: 1, 176
89: 1, 17563: 1, 17380: 1, 17371: 1, 17290: 1, 17158: 1, 17130: 1, 17057: 1, 16930: 1, 16884: 1, 1
6497: 1, 16332: 1, 16129: 1, 15846: 1, 15732: 1, 15706: 1, 15631: 1, 15607: 1, 15584: 1, 15516: 1,
15482: 1, 15412: 1, 15391: 1, 15330: 1, 15089: 1, 14989: 1, 14860: 1, 14780: 1, 14701: 1, 14594: 1
, 14520: 1, 14399: 1, 14338: 1, 14249: 1, 14215: 1, 14176: 1, 14130: 1, 13754: 1, 13678: 1, 13660:
1, 13580: 1, 13428: 1, 13229: 1, 13220: 1, 13185: 1, 13143: 1, 13135: 1, 13053: 1, 13042: 1, 13002
: 1, 13000: 1, 12886: 1, 12879: 1, 12784: 1, 12707: 1, 12632: 1, 12627: 1, 12588: 1, 12577: 1, 125
50: 1, 12366: 1, 12349: 1, 12347: 1, 12259: 1, 12256: 1, 12227: 1, 12198: 1, 12193: 1, 12188: 1, 1
2179: 1, 12133: 1, 12103: 1, 12056: 1, 11991: 1, 11950: 1, 11941: 1, 11928: 1, 11880: 1, 11801: 1,
11795: 1, 11766: 1, 11664: 1, 11645: 1, 11610: 1, 11580: 1, 11571: 1, 11460: 1, 11457: 1, 11446: 1
, 11357: 1, 11269: 1, 11208: 1, 11180: 1, 11170: 1, 11063: 1, 11025: 1, 11018: 1, 11012: 1, 10851:
1, 10811: 1, 10770: 1, 10674: 1, 10518: 1, 10479: 1, 10445: 1, 10406: 1, 10366: 1, 10353: 1, 10342
: 1, 10318: 1, 10253: 1, 10141: 1, 10092: 1, 10080: 1, 9980: 1, 9962: 1, 9960: 1, 9944: 1, 9889: 1,
9886: 1, 9877: 1, 9869: 1, 9852: 1, 9847: 1, 9813: 1, 9777: 1, 9773: 1, 9766: 1, 9589: 1, 9441: 1,
9429: 1, 9408: 1, 9371: 1, 9346: 1, 9317: 1, 9287: 1, 9273: 1, 9233: 1, 9231: 1, 9200: 1, 9191: 1,
9154: 1, 9143: 1, 9128: 1, 9121: 1, 9062: 1, 9019: 1, 9008: 1, 9004: 1, 9002: 1, 8999: 1, 8918: 1,
8914: 1, 8886: 1, 8849: 1, 8842: 1, 8840: 1, 8803: 1, 8722: 1, 8718: 1, 8701: 1, 8613: 1, 8594: 1,
8569: 1, 8545: 1, 8541: 1, 8531: 1, 8472: 1, 8447: 1, 8437: 1, 8342: 1, 8329: 1, 8327: 1, 8325: 1,
8308: 1, 8306: 1, 8250: 1, 8243: 1, 8235: 1, 8225: 1, 8178: 1, 8164: 1, 8142: 1, 8126: 1, 8107: 1,
8099: 1, 8086: 1, 8065: 1, 8052: 1, 8051: 1, 8036: 1, 8035: 1, 7977: 1, 7930: 1, 7896: 1, 7821: 1,
7808: 1, 7804: 1, 7776: 1, 7775: 1, 7756: 1, 7686: 1, 7643: 1, 7617: 1, 7597: 1, 7593: 1, 7569: 1,
7504: 1, 7497: 1, 7474: 1, 7456: 1, 7443: 1, 7440: 1, 7420: 1, 7338: 1, 7332: 1, 7330: 1, 7309: 1,
7306: 1, 7305: 1, 7285: 1, 7258: 1, 7236: 1, 7225: 1, 7207: 1, 7203: 1, 7200: 1, 7179: 1, 7151: 1,
7143: 1, 7129: 1, 7118: 1, 7095: 1, 7083: 1, 7081: 1, 7073: 1, 7065: 1, 7043: 1, 7012: 1, 6987: 1,
6982: 1, 6980: 1, 6977: 1, 6968: 1, 6954: 1, 6950: 1, 6906: 1, 6897: 1, 6890: 1, 6873: 1, 6863: 1,
6858: 1, 6857: 1, 6810: 1, 6803: 1, 6786: 1, 6772: 1, 6765: 1, 6732: 1, 6730: 1, 6727: 1, 6711: 1,
6694: 1, 6693: 1, 6657: 1, 6603: 1, 6589: 1, 6584: 1, 6578: 1, 6549: 1, 6520: 1, 6504: 1, 6499: 1,
6474: 1, 6470: 1, 6468: 1, 6438: 1, 6436: 1, 6422: 1, 6396: 1, 6379: 1, 6357: 1, 6321: 1, 6320: 1,
6303: 1, 6294: 1, 6293: 1, 6291: 1, 6290: 1, 6274: 1, 6259: 1, 6252: 1, 6233: 1, 6231: 1, 6219: 1,
6161: 1, 6141: 1, 6111: 1, 6102: 1, 6082: 1, 6071: 1, 6055: 1, 6040: 1, 6039: 1, 6025: 1, 6024: 1,
6017: 1, 6008: 1, 6001: 1, 5965: 1, 5927: 1, 5921: 1, 5907: 1, 5887: 1, 5878: 1, 5865: 1, 5864: 1,
5852: 1, 5812: 1, 5809: 1, 5780: 1, 5778: 1, 5775: 1, 5750: 1, 5742: 1, 5734: 1, 5730: 1, 5711: 1,
5678: 1, 5647: 1, 5632: 1, 5625: 1, 5611: 1, 5602: 1, 5582: 1, 5575: 1, 5538: 1, 5531: 1, 5506: 1,
5483: 1, 5475: 1, 5468: 1, 5453: 1, 5443: 1, 5436: 1, 5430: 1, 5426: 1, 5410: 1, 5405: 1, 5404: 1,
5398: 1, 5396: 1, 5352: 1, 5336: 1, 5326: 1, 5291: 1, 5262: 1, 5259: 1, 5245: 1, 5223: 1, 5222: 1,
5217: 1, 5209: 1, 5192: 1, 5172: 1, 5171: 1, 5166: 1, 5150: 1, 5137: 1, 5136: 1, 5129: 1, 5112: 1,
5109: 1, 5085: 1, 5068: 1, 5055: 1, 5044: 1, 5037: 1, 5035: 1, 5022: 1, 5009: 1, 5008: 1, 4997: 1,
4993: 1, 4992: 1, 4976: 1, 4969: 1, 4966: 1, 4950: 1, 4944: 1, 4940: 1, 4929: 1, 4925: 1, 4921: 1,
4910: 1, 4905: 1, 4883: 1, 4874: 1, 4861: 1, 4854: 1, 4852: 1, 4846: 1, 4838: 1, 4835: 1, 4834: 1,
4829: 1, 4801: 1, 4796: 1, 4784: 1, 4782: 1, 4780: 1, 4779: 1, 4772: 1, 4771: 1, 4760: 1, 4756: 1,
4746: 1, 4717: 1, 4710: 1, 4697: 1, 4671: 1, 4664: 1, 4659: 1, 4651: 1, 4639: 1, 4618: 1, 4613: 1,
4596: 1, 4588: 1, 4585: 1, 4575: 1, 4567: 1, 4555: 1, 4550: 1, 4525: 1, 4503: 1, 4498: 1, 4487: 1,
4486: 1, 4482: 1, 4477: 1, 4466: 1, 4464: 1, 4447: 1, 4431: 1, 4418: 1, 4416: 1, 4415: 1, 4394: 1,
4392: 1, 4386: 1, 4372: 1, 4371: 1, 4352: 1, 4342: 1, 4340: 1, 4338: 1, 4321: 1, 4309: 1, 4285: 1,
4283: 1, 4282: 1, 4273: 1, 4271: 1, 4267: 1, 4263: 1, 4257: 1, 4252: 1, 4244: 1, 4236: 1, 4229: 1,
4228: 1, 4226: 1, 4222: 1, 4220: 1, 4218: 1, 4217: 1, 4215: 1, 4204: 1, 4199: 1, 4196: 1, 4192: 1,
4175: 1, 4172: 1, 4168: 1, 4159: 1, 4149: 1, 4146: 1, 4144: 1, 4130: 1, 4123: 1, 4111: 1, 4106: 1,
4089: 1, 4074: 1, 4066: 1, 4061: 1, 4060: 1, 4056: 1, 4055: 1, 4042: 1, 4030: 1, 4025: 1, 4021: 1,
4010: 1, 4005: 1, 4001: 1, 3998: 1, 3995: 1, 3971: 1, 3967: 1, 3965: 1, 3960: 1, 3947: 1, 3943: 1,
3941: 1, 3934: 1, 3908: 1, 3901: 1, 3892: 1, 3865: 1, 3856: 1, 3846: 1, 3832: 1, 3828: 1, 3821: 1,
3820: 1, 3819: 1, 3817: 1, 3816: 1, 3809: 1, 3804: 1, 3798: 1, 3796: 1, 3787: 1, 3778: 1, 3769: 1,
3768: 1, 3760: 1, 3759: 1, 3754: 1, 3752: 1, 3742: 1, 3740: 1, 3739: 1, 3738: 1, 3736: 1, 3731: 1,
3720: 1, 3717: 1, 3711: 1, 3709: 1, 3706: 1, 3702: 1, 3699: 1, 3697: 1, 3686: 1, 3682: 1, 3676: 1,
3666: 1, 3653: 1, 3650: 1, 3647: 1, 3646: 1, 3645: 1, 3639: 1, 3628: 1, 3615: 1, 3610: 1, 3603: 1,
3602: 1, 3590: 1, 3588: 1, 3587: 1, 3576: 1, 3573: 1, 3566: 1, 3560: 1, 3554: 1, 3545: 1, 3543: 1,
3542: 1, 3541: 1, 3531: 1, 3523: 1, 3522: 1, 3521: 1, 3518: 1, 3517: 1, 3512: 1, 3500: 1, 3494: 1,
3486: 1, 3480: 1, 3478: 1, 3469: 1, 3467: 1, 3466: 1, 3465: 1, 3461: 1, 3454: 1, 3451: 1, 3448: 1,
3445: 1, 3441: 1, 3436: 1, 3434: 1, 3433: 1, 3413: 1, 3411: 1, 3406: 1, 3403: 1, 3401: 1, 3400: 1,
3397: 1, 3396: 1, 3383: 1, 3381: 1, 3380: 1, 3379: 1, 3378: 1, 3369: 1, 3359: 1, 3356: 1, 3354: 1,
3345: 1, 3344: 1, 3342: 1, 3339: 1, 3333: 1, 3332: 1, 3327: 1, 3317: 1, 3314: 1, 3308: 1, 3305: 1,
3288: 1, 3284: 1, 3282: 1, 3275: 1, 3272: 1, 3268: 1, 3264: 1, 3263: 1, 3245: 1, 3242: 1, 3239: 1,
3235: 1, 3228: 1, 3224: 1, 3221: 1, 3212: 1, 3211: 1, 3205: 1, 3203: 1, 3192: 1, 3185: 1, 3182: 1,
3178: 1, 3177: 1, 3176: 1, 3171: 1, 3167: 1, 3161: 1, 3160: 1, 3157: 1, 3153: 1, 3144: 1, 3143: 1,
3139: 1, 3136: 1, 3122: 1, 3116: 1, 3114: 1, 3101: 1, 3093: 1, 3089: 1, 3083: 1, 3077: 1, 3072: 1,
3068: 1, 3063: 1, 3056: 1, 3054: 1, 3048: 1, 3042: 1, 3038: 1, 3024: 1, 3015: 1, 3013: 1, 3010: 1,
3006: 1, 3002: 1, 3001: 1, 2992: 1, 2986: 1, 2978: 1, 2974: 1, 2970: 1, 2965: 1, 2964: 1, 2957: 1,
2956: 1, 2953: 1, 2951: 1, 2950: 1, 2946: 1, 2942: 1, 2941: 1, 2928: 1, 2926: 1, 2921: 1, 2918: 1,
2914: 1, 2912: 1, 2911: 1, 2909: 1, 2903: 1, 2902: 1, 2901: 1, 2896: 1, 2892: 1, 2884: 1, 2872: 1,
2871: 1, 2866: 1, 2863: 1, 2861: 1, 2858: 1, 2856: 1, 2849: 1, 2840: 1, 2832: 1, 2829: 1, 2826: 1,
2818: 1, 2815: 1, 2814: 1, 2810: 1, 2807: 1, 2797: 1, 2785: 1, 2784: 1, 2780: 1, 2778: 1, 2771: 1,

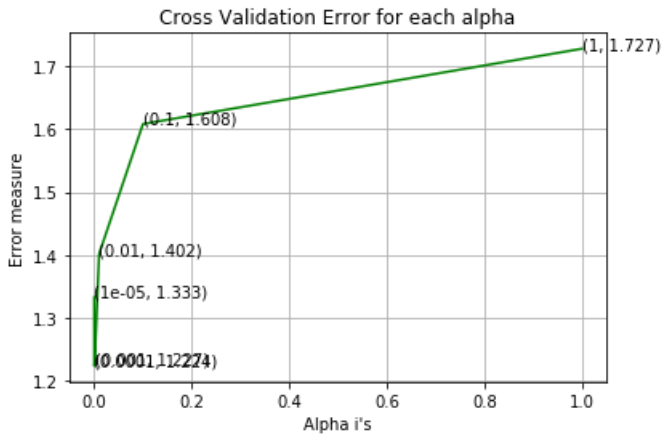
2818: 1, 2815: 1, 2814: 1, 2810: 1, 2807: 1, 2797: 1, 2785: 1, 2784: 1, 2780: 1, 2778: 1, 2771: 1,
2767: 1, 2760: 1, 2758: 1, 2752: 1, 2746: 1, 2742: 1, 2741: 1, 2733: 1, 2722: 1, 2703: 1, 2691: 1,
2690: 1, 2688: 1, 2686: 1, 2682: 1, 2678: 1, 2677: 1, 2676: 1, 2673: 1, 2672: 1, 2668: 1, 2666: 1,
2653: 1, 2645: 1, 2638: 1, 2637: 1, 2628: 1, 2620: 1, 2617: 1, 2616: 1, 2614: 1, 2602: 1, 2599: 1,
2588: 1, 2579: 1, 2574: 1, 2568: 1, 2566: 1, 2560: 1, 2552: 1, 2549: 1, 2544: 1, 2542: 1, 2537: 1,
2531: 1, 2530: 1, 2529: 1, 2526: 1, 2525: 1, 2523: 1, 2522: 1, 2517: 1, 2508: 1, 2506: 1, 2497: 1,
2486: 1, 2484: 1, 2483: 1, 2476: 1, 2473: 1, 2470: 1, 2462: 1, 2460: 1, 2458: 1, 2456: 1, 2454: 1,
2453: 1, 2452: 1, 2451: 1, 2449: 1, 2444: 1, 2441: 1, 2436: 1, 2430: 1, 2424: 1, 2418: 1, 2416: 1,
2412: 1, 2411: 1, 2410: 1, 2408: 1, 2407: 1, 2405: 1, 2398: 1, 2397: 1, 2393: 1, 2389: 1, 2387: 1,
2386: 1, 2385: 1, 2377: 1, 2374: 1, 2373: 1, 2371: 1, 2362: 1, 2361: 1, 2357: 1, 2355: 1, 2353: 1,
2352: 1, 2347: 1, 2346: 1, 2344: 1, 2343: 1, 2341: 1, 2340: 1, 2334: 1, 2329: 1, 2324: 1, 2321: 1,
2320: 1, 2319: 1, 2318: 1, 2317: 1, 2315: 1, 2306: 1, 2302: 1, 2301: 1, 2300: 1, 2291: 1, 2288: 1,
2287: 1, 2284: 1, 2280: 1, 2274: 1, 2273: 1, 2272: 1, 2263: 1, 2261: 1, 2257: 1, 2254: 1, 2247: 1,
2246: 1, 2245: 1, 2241: 1, 2238: 1, 2225: 1, 2219: 1, 2218: 1, 2207: 1, 2203: 1, 2202: 1, 2200: 1,
2199: 1, 2195: 1, 2192: 1, 2189: 1, 2187: 1, 2186: 1, 2184: 1, 2181: 1, 2180: 1, 2178: 1, 2177: 1,
2175: 1, 2172: 1, 2168: 1, 2159: 1, 2158: 1, 2156: 1, 2155: 1, 2153: 1, 2150: 1, 2141: 1, 2138: 1,
2137: 1, 2132: 1, 2129: 1, 2124: 1, 2115: 1, 2113: 1, 2108: 1, 2103: 1, 2100: 1, 2089: 1, 2085: 1,
2083: 1, 2080: 1, 2079: 1, 2077: 1, 2070: 1, 2068: 1, 2063: 1, 2062: 1, 2059: 1, 2058: 1, 2057: 1,
2056: 1, 2055: 1, 2053: 1, 2050: 1, 2049: 1, 2044: 1, 2040: 1, 2039: 1, 2038: 1, 2035: 1, 2030: 1,
2029: 1, 2026: 1, 2024: 1, 2023: 1, 2022: 1, 2018: 1, 2016: 1, 2011: 1, 2006: 1, 2005: 1, 2000: 1,
1999: 1, 1996: 1, 1995: 1, 1994: 1, 1992: 1, 1986: 1, 1985: 1, 1983: 1, 1980: 1, 1978: 1, 1972: 1,
1969: 1, 1967: 1, 1963: 1, 1962: 1, 1952: 1, 1947: 1, 1946: 1, 1941: 1, 1939: 1, 1937: 1, 1932: 1,
1931: 1, 1929: 1, 1926: 1, 1922: 1, 1920: 1, 1919: 1, 1911: 1, 1910: 1, 1905: 1, 1904: 1, 1900: 1,
1899: 1, 1895: 1, 1888: 1, 1885: 1, 1883: 1, 1881: 1, 1874: 1, 1871: 1, 1870: 1, 1867: 1, 1864: 1,
1863: 1, 1862: 1, 1858: 1, 1856: 1, 1855: 1, 1852: 1, 1849: 1, 1846: 1, 1842: 1, 1838: 1, 1833: 1,
1830: 1, 1829: 1, 1827: 1, 1824: 1, 1823: 1, 1820: 1, 1817: 1, 1813: 1, 1811: 1, 1807: 1, 1794: 1,
1791: 1, 1790: 1, 1788: 1, 1783: 1, 1781: 1, 1776: 1, 1775: 1, 1773: 1, 1770: 1, 1768: 1, 1765: 1,
1761: 1, 1759: 1, 1758: 1, 1755: 1, 1753: 1, 1751: 1, 1750: 1, 1749: 1, 1746: 1, 1744: 1, 1743: 1,
1742: 1, 1739: 1, 1738: 1, 1736: 1, 1735: 1, 1731: 1, 1728: 1, 1726: 1, 1723: 1, 1721: 1, 1720: 1,
1714: 1, 1709: 1, 1706: 1, 1705: 1, 1704: 1, 1701: 1, 1700: 1, 1689: 1, 1686: 1, 1683: 1, 1679: 1,
1675: 1, 1672: 1, 1665: 1, 1664: 1, 1662: 1, 1655: 1, 1653: 1, 1648: 1, 1647: 1, 1643: 1, 1641: 1,
1638: 1, 1637: 1, 1635: 1, 1633: 1, 1632: 1, 1630: 1, 1624: 1, 1619: 1, 1618: 1, 1616: 1, 1614: 1,
1608: 1, 1605: 1, 1603: 1, 1602: 1, 1600: 1, 1597: 1, 1595: 1, 1594: 1, 1593: 1, 1586: 1, 1585: 1,
1584: 1, 1582: 1, 1581: 1, 1580: 1, 1579: 1, 1578: 1, 1577: 1, 1576: 1, 1575: 1, 1573: 1, 1564: 1,
1560: 1, 1555: 1, 1554: 1, 1549: 1, 1538: 1, 1537: 1, 1526: 1, 1525: 1, 1524: 1, 1522: 1, 1519: 1,
1517: 1, 1516: 1, 1511: 1, 1510: 1, 1507: 1, 1506: 1, 1505: 1, 1504: 1, 1503: 1, 1501: 1, 1498: 1,
1492: 1, 1491: 1, 1489: 1, 1485: 1, 1484: 1, 1481: 1, 1479: 1, 1475: 1, 1474: 1, 1472: 1, 1470: 1,
1468: 1, 1461: 1, 1460: 1, 1458: 1, 1457: 1, 1456: 1, 1455: 1, 1454: 1, 1449: 1, 1448: 1, 1446: 1,
1445: 1, 1443: 1, 1442: 1, 1435: 1, 1433: 1, 1432: 1, 1431: 1, 1429: 1, 1427: 1, 1423: 1, 1419: 1,
1418: 1, 1417: 1, 1416: 1, 1414: 1, 1410: 1, 1403: 1, 1402: 1, 1401: 1, 1400: 1, 1394: 1, 1386: 1,
1384: 1, 1383: 1, 1375: 1, 1373: 1, 1369: 1, 1364: 1, 1363: 1, 1359: 1, 1354: 1, 1353: 1, 1351: 1,
1348: 1, 1346: 1, 1341: 1, 1340: 1, 1339: 1, 1334: 1, 1331: 1, 1328: 1, 1326: 1, 1325: 1, 1324: 1,
1321: 1, 1320: 1, 1319: 1, 1318: 1, 1315: 1, 1313: 1, 1311: 1, 1305: 1, 1302: 1, 1301: 1, 1299: 1,
1296: 1, 1292: 1, 1286: 1, 1280: 1, 1278: 1, 1273: 1, 1271: 1, 1268: 1, 1260: 1, 1258: 1, 1254: 1,
1253: 1, 1252: 1, 1246: 1, 1245: 1, 1243: 1, 1236: 1, 1234: 1, 1233: 1, 1231: 1, 1229: 1, 1228: 1,
1226: 1, 1225: 1, 1221: 1, 1218: 1, 1217: 1, 1212: 1, 1209: 1, 1207: 1, 1198: 1, 1197: 1, 1190: 1,
1188: 1, 1185: 1, 1183: 1, 1178: 1, 1175: 1, 1173: 1, 1171: 1, 1170: 1, 1169: 1, 1166: 1, 1164: 1,
1163: 1, 1162: 1, 1158: 1, 1155: 1, 1150: 1, 1146: 1, 1144: 1, 1143: 1, 1140: 1, 1139: 1, 1138: 1,
1137: 1, 1135: 1, 1131: 1, 1125: 1, 1124: 1, 1122: 1, 1117: 1, 1115: 1, 1114: 1, 1113: 1, 1110: 1,
1108: 1, 1107: 1, 1102: 1, 1099: 1, 1094: 1, 1090: 1, 1088: 1, 1084: 1, 1083: 1, 1080: 1, 1078: 1,
1074: 1, 1072: 1, 1070: 1, 1068: 1, 1065: 1, 1057: 1, 1052: 1, 1050: 1, 1049: 1, 1046: 1, 1044: 1,
1041: 1, 1039: 1, 1031: 1, 1030: 1, 1029: 1, 1027: 1, 1026: 1, 1025: 1, 1024: 1, 1019: 1, 1013: 1,
1006: 1, 1005: 1, 1002: 1, 1000: 1, 998: 1, 997: 1, 992: 1, 988: 1, 987: 1, 986: 1, 976: 1, 973: 1,
969: 1, 968: 1, 964: 1, 963: 1, 962: 1, 960: 1, 956: 1, 952: 1, 950: 1, 947: 1, 945: 1, 942: 1,
935: 1, 930: 1, 920: 1, 918: 1, 917: 1, 911: 1, 906: 1, 900: 1, 899: 1, 898: 1, 897: 1, 893: 1,
892: 1, 891: 1, 887: 1, 882: 1, 881: 1, 878: 1, 877: 1, 874: 1, 862: 1, 859: 1, 858: 1, 853: 1,
852: 1, 848: 1, 832: 1, 830: 1, 816: 1, 805: 1, 796: 1, 794: 1, 793: 1, 785: 1, 782: 1, 780: 1,
766: 1, 765: 1, 760: 1, 757: 1, 750: 1, 743: 1, 740: 1, 727: 1, 724: 1, 718: 1, 715: 1, 712: 1,
707: 1, 705: 1, 700: 1, 691: 1, 660: 1, 657: 1, 654: 1, 636: 1, 626: 1, 624: 1, 622: 1, 598: 1,
593: 1, 591: 1, 571: 1, 545: 1, 526: 1, 520: 1, 489: 1, 487: 1, 478: 1, 357: 1})

In [0]:

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

# read more about SGDClassifier() at http://scikit-
learn.org/stable/modules/generated/sklearn.linear_model.SGDClassifier.html
# ------------------------------
# default parameters
# SGDClassifier(loss='hinge', penalty='l2', alpha=0.0001, l1_ratio=0.15, fit_intercept=True, max_i
ter=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
```

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

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


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

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

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


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

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

```
For values of alpha =  1e-05 The log loss is: 1.3330368917617879
For values of alpha =  0.0001 The log loss is: 1.2241898008847032
For values of alpha =  0.001 The log loss is: 1.227115984872649
For values of alpha =  0.01 The log loss is: 1.4016584121734113
For values of alpha =  0.1 The log loss is: 1.6078903681851084
For values of alpha =  1 The log loss is: 1.7272690445029124
```




```
For values of best alpha =  0.0001 The train log loss is: 0.7349137188354572
For values of best alpha =  0.0001 The cross validation log loss is: 1.2241898008847032
For values of best alpha =  0.0001 The test log loss is: 1.1553032120657891
```

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 = CountVectorizer(min_df=3)
    df_text_fea = df_text_vec.fit_transform(df['TEXT'])
    df_text_features = df_text_vec.get_feature_names()

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

In [0]:

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

```
96.918 % of word of test data appeared in train data
98.498 % 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 probabilities belongs to each class
    print("Log loss :",log_loss(test_y, sig_clf.predict_proba(test_x)))
    # calculating the number of data points that are misclassified
    print("Number of mis-classified points :", np.count_nonzero((pred_y- test_y))/test_y.shape[0])
    plot_confusion_matrix(test_y, pred_y)
```

In [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 = CountVectorizer()
    var_count_vec = CountVectorizer()
    text_count_vec = CountVectorizer(min_df=3)

    gene_vec = gene_count_vec.fit(train_df['Gene'])
```

```
    var_vec  = var_count_vec.fit(train_df['Variation'])
    text_vec = text_count_vec.fit(train_df['TEXT'])

    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_r
o))
        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)).tocs
r()
train_y = np.array(list(train_df['Class']))

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

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


train_gene_var_responseCoding =
np.hstack((train_gene_feature_responseCoding,train_variation_feature_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 [0]:

```
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, 57039)
(number of data points * number of features) in test data =  (665, 57039)
(number of data points * number of features) in cross validation data = (532, 57039)
```

In [0]:

```
print(" Response encoding features :")
print("(number of data points * number of features) in train data = ", train_x_responseCoding.shap
e)
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)
```

4.1. Base Line Model

4.1.1. Naive Bayes

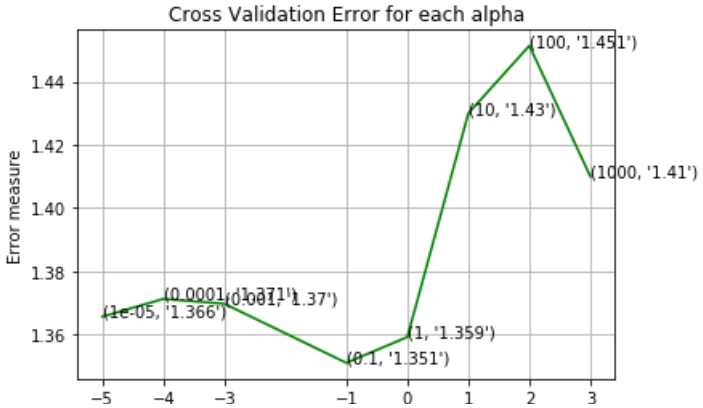
4.1.1.1. Hyper parameter tuning

In [0]:

```
# 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, method='sigmoid', cv=3)
#
# some of the methods of CalibratedClassifierCV()
# fit(X, y[, sample_weight]) Fit the calibrated model
# get_params([deep]) Get parameters for this estimator.
# predict(X) Predict the target of new samples.
# predict_proba(X) Posterior probabilities of classification
# ----------------------------
# video link: 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_array[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 validation log loss is:",log_lo
ss(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, p
redict_y, labels=clf.classes_, eps=1e-15))
```

```
for alpha = 1e-05
Log Loss : 1.3655537837941127
for alpha = 0.0001
Log Loss : 1.3711793514758466
for alpha = 0.001
Log Loss : 1.3696542195047903
for alpha = 0.1
Log Loss : 1.3509235125982695
for alpha = 1
Log Loss : 1.3591155403248767
for alpha = 10
Log Loss : 1.4299766791532638
for alpha = 100
Log Loss : 1.451360452876549
for alpha = 1000
Log Loss : 1.4099515277732073
```



Alpha i's

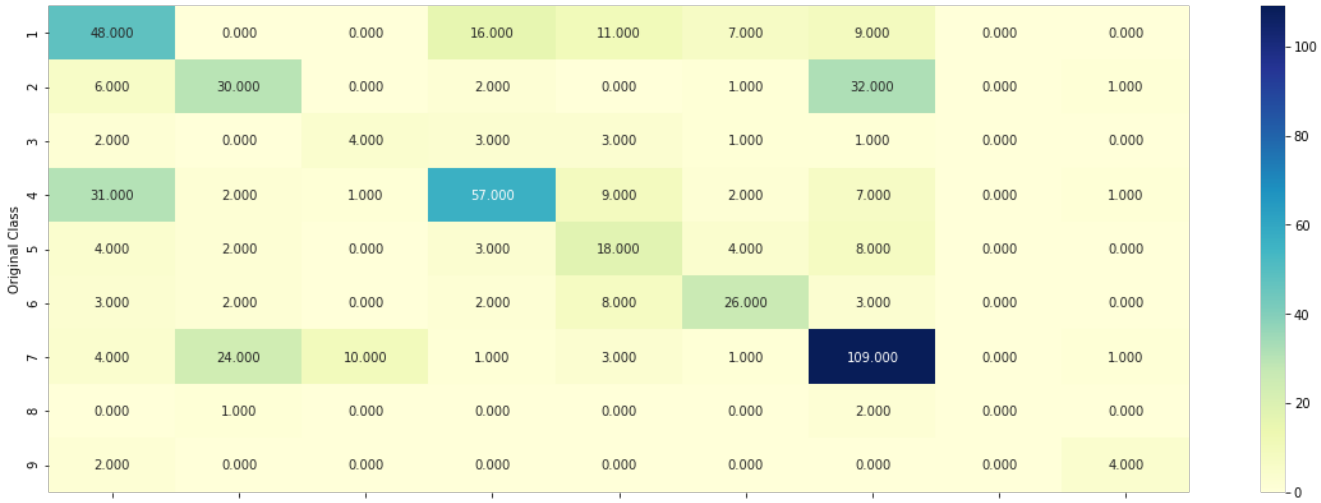
```
For values of best alpha =  0.1 The train log loss is: 0.9033118479519152
For values of best alpha =  0.1 The cross validation log loss is: 1.3509235125982695
For values of best alpha =  0.1 The test log loss is: 1.2784897724659714
```

4.1.1.2. Testing the model with best hyper paramters

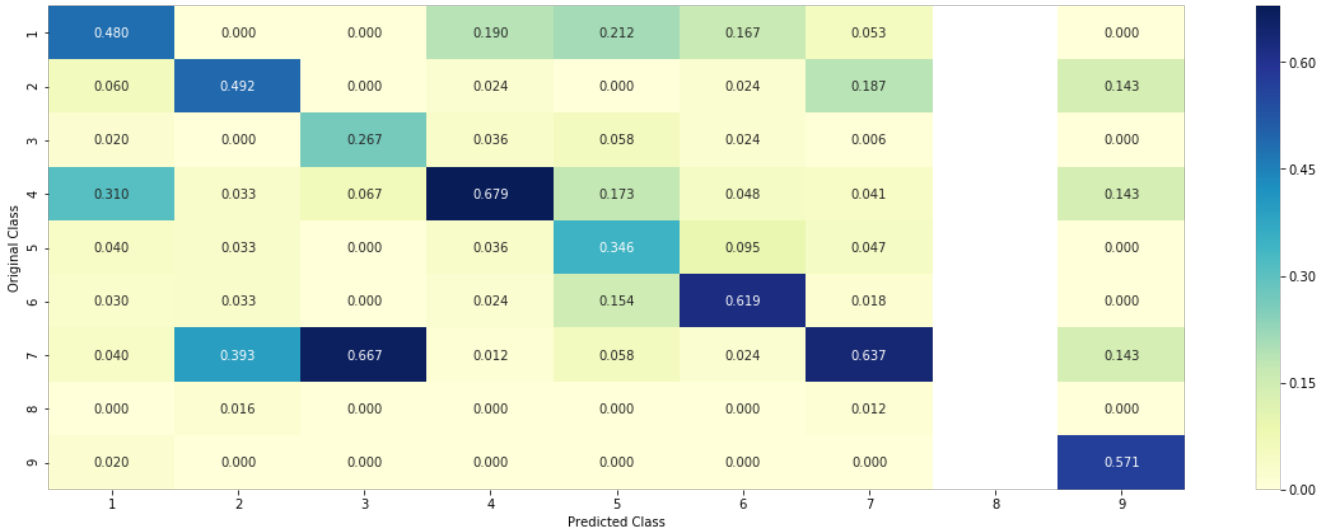
In [0]:

```
# 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, method='sigmoid', cv=3)
#
# some of the methods of CalibratedClassifierCV()
# fit(X, y[, sample_weight]) Fit the calibrated model
# get_params([deep]) Get parameters for this estimator.
# predict(X) Predict the target of new samples.
# predict_proba(X) Posterior probabilities of classification
# ----------------------------

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-probability estimates
print("Log Loss :",log_loss(cv_y, sig_clf_probs))
print("Number of missclassified point :", np.count_nonzero((sig_clf.predict(cv_x_onehotCoding)- cv
_y))/cv_y.shape[0])
plot_confusion_matrix(cv_y, sig_clf.predict(cv_x_onehotCoding.toarray()))
```

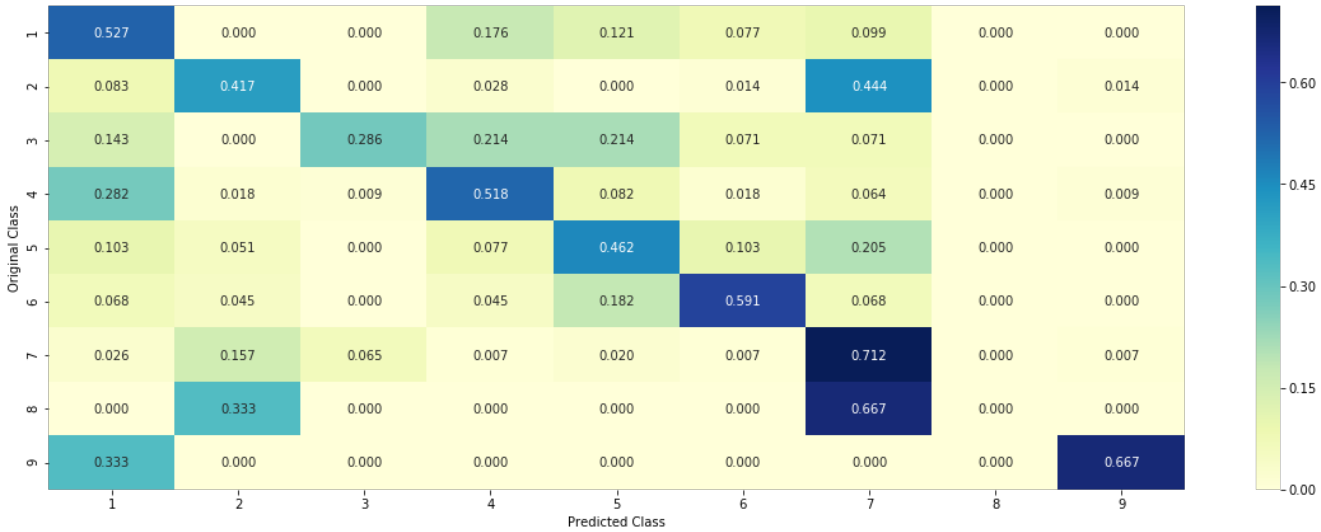
```
Log Loss : 1.3509235125982695
Number of missclassified point : 0.44360902255639095
-------------------- Confusion matrix --------------------
```




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




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



4.1.1.3. Feature Importance, Correctly classified point

In [0]:

```
test_point_index = 1
no_feature = 100
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],
test_df['TEXT'].iloc[test_point_index],test_df['Gene'].iloc[test_point_index],test_df['Variation']
.iloc[test_point_index], no_feature)
```

```
Predicted Class : 7
Predicted Class Probabilities: [[0.0911 0.1284 0.0141 0.1119 0.0334 0.0365 0.5764 0.0057 0.0025]]
Actual Class : 7
--------------------------------------------------
16 Text feature [presence] present in test data point [True]
17 Text feature [kinase] present in test data point [True]
```

18 Text feature [well] present in test data point [True]
19 Text feature [activating] present in test data point [True]
20 Text feature [downstream] present in test data point [True]
21 Text feature [cell] present in test data point [True]
22 Text feature [inhibitor] present in test data point [True]
23 Text feature [cells] present in test data point [True]
24 Text feature [independent] present in test data point [True]
25 Text feature [contrast] present in test data point [True]
26 Text feature [recently] present in test data point [True]
29 Text feature [shown] present in test data point [True]
30 Text feature [potential] present in test data point [True]
31 Text feature [also] present in test data point [True]
32 Text feature [obtained] present in test data point [True]
33 Text feature [growth] present in test data point [True]
34 Text feature [activation] present in test data point [True]
35 Text feature [suggest] present in test data point [True]
36 Text feature [showed] present in test data point [True]
37 Text feature [however] present in test data point [True]
38 Text feature [expressing] present in test data point [True]
39 Text feature [addition] present in test data point [True]
40 Text feature [found] present in test data point [True]
41 Text feature [10] present in test data point [True]
42 Text feature [previously] present in test data point [True]
43 Text feature [factor] present in test data point [True]
44 Text feature [compared] present in test data point [True]
45 Text feature [treated] present in test data point [True]
46 Text feature [inhibition] present in test data point [True]
47 Text feature [higher] present in test data point [True]
48 Text feature [observed] present in test data point [True]
49 Text feature [described] present in test data point [True]
50 Text feature [may] present in test data point [True]
51 Text feature [similar] present in test data point [True]
52 Text feature [total] present in test data point [True]
53 Text feature [furthermore] present in test data point [True]
54 Text feature [studies] present in test data point [True]
55 Text feature [using] present in test data point [True]
56 Text feature [without] present in test data point [True]
57 Text feature [concentrations] present in test data point [True]
58 Text feature [1a] present in test data point [True]
59 Text feature [various] present in test data point [True]
60 Text feature [including] present in test data point [True]
61 Text feature [mutations] present in test data point [True]
62 Text feature [respectively] present in test data point [True]
63 Text feature [12] present in test data point [True]
64 Text feature [followed] present in test data point [True]
65 Text feature [enhanced] present in test data point [True]
66 Text feature [although] present in test data point [True]
67 Text feature [interestingly] present in test data point [True]
68 Text feature [phosphorylation] present in test data point [True]
70 Text feature [new] present in test data point [True]
71 Text feature [inhibited] present in test data point [True]
72 Text feature [constitutively] present in test data point [True]
75 Text feature [1b] present in test data point [True]
76 Text feature [reported] present in test data point [True]
77 Text feature [confirmed] present in test data point [True]
78 Text feature [inhibitors] present in test data point [True]
79 Text feature [proliferation] present in test data point [True]
80 Text feature [report] present in test data point [True]
81 Text feature [either] present in test data point [True]
82 Text feature [molecular] present in test data point [True]
83 Text feature [15] present in test data point [True]
84 Text feature [thus] present in test data point [True]
85 Text feature [recent] present in test data point [True]
86 Text feature [3b] present in test data point [True]
87 Text feature [fig] present in test data point [True]
88 Text feature [results] present in test data point [True]
89 Text feature [occur] present in test data point [True]
90 Text feature [small] present in test data point [True]
91 Text feature [3a] present in test data point [True]
92 Text feature [approximately] present in test data point [True]
93 Text feature [hours] present in test data point [True]
94 Text feature [consistent] present in test data point [True]
95 Text feature [figure] present in test data point [True]
97 Text feature [suggests] present in test data point [True]
98 Text feature [absence] present in test data point [True]
99 Text feature [measured] present in test data point [True]
Out of the top 100 features 78 are present in query point

4.1.1.4. Feature Importance, Incorrectly classified point

In [0]:

```
test_point_index = 100
no_feature = 100
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],
test_df['TEXT'].iloc[test_point_index],test_df['Gene'].iloc[test_point_index],test_df['Variation']
.iloc[test_point_index], no_feature)
```

```
Predicted Class : 7
Predicted Class Probabilities: [[0.0876 0.0895 0.0136 0.1069 0.0322 0.0356 0.6267 0.0055 0.0024]]
Actual Class : 7
--------------------------------------------------
17 Text feature [kinase] present in test data point [True]
18 Text feature [well] present in test data point [True]
19 Text feature [activating] present in test data point [True]
20 Text feature [downstream] present in test data point [True]
21 Text feature [cell] present in test data point [True]
23 Text feature [cells] present in test data point [True]
24 Text feature [independent] present in test data point [True]
25 Text feature [contrast] present in test data point [True]
26 Text feature [recently] present in test data point [True]
29 Text feature [shown] present in test data point [True]
30 Text feature [potential] present in test data point [True]
31 Text feature [also] present in test data point [True]
33 Text feature [growth] present in test data point [True]
34 Text feature [activation] present in test data point [True]
35 Text feature [suggest] present in test data point [True]
36 Text feature [showed] present in test data point [True]
37 Text feature [however] present in test data point [True]
39 Text feature [addition] present in test data point [True]
40 Text feature [found] present in test data point [True]
41 Text feature [10] present in test data point [True]
42 Text feature [previously] present in test data point [True]
44 Text feature [compared] present in test data point [True]
46 Text feature [inhibition] present in test data point [True]
47 Text feature [higher] present in test data point [True]
48 Text feature [observed] present in test data point [True]
50 Text feature [may] present in test data point [True]
51 Text feature [similar] present in test data point [True]
54 Text feature [studies] present in test data point [True]
55 Text feature [using] present in test data point [True]
56 Text feature [without] present in test data point [True]
58 Text feature [1a] present in test data point [True]
60 Text feature [including] present in test data point [True]
61 Text feature [mutations] present in test data point [True]
62 Text feature [respectively] present in test data point [True]
63 Text feature [12] present in test data point [True]
64 Text feature [followed] present in test data point [True]
65 Text feature [enhanced] present in test data point [True]
66 Text feature [although] present in test data point [True]
68 Text feature [phosphorylation] present in test data point [True]
69 Text feature [activated] present in test data point [True]
70 Text feature [new] present in test data point [True]
71 Text feature [inhibited] present in test data point [True]
72 Text feature [constitutively] present in test data point [True]
75 Text feature [1b] present in test data point [True]
78 Text feature [inhibitors] present in test data point [True]
79 Text feature [proliferation] present in test data point [True]
81 Text feature [either] present in test data point [True]
82 Text feature [molecular] present in test data point [True]
83 Text feature [15] present in test data point [True]
85 Text feature [recent] present in test data point [True]
86 Text feature [3b] present in test data point [True]
87 Text feature [fig] present in test data point [True]
```

```
88 Text feature [results] present in test data point [True]
89 Text feature [occur] present in test data point [True]
90 Text feature [small] present in test data point [True]
91 Text feature [3a] present in test data point [True]
94 Text feature [consistent] present in test data point [True]
95 Text feature [figure] present in test data point [True]
97 Text feature [suggests] present in test data point [True]
98 Text feature [absence] present in test data point [True]
Out of the top  100  features  60 are present in query point
```

4.2. K Nearest Neighbour Classification

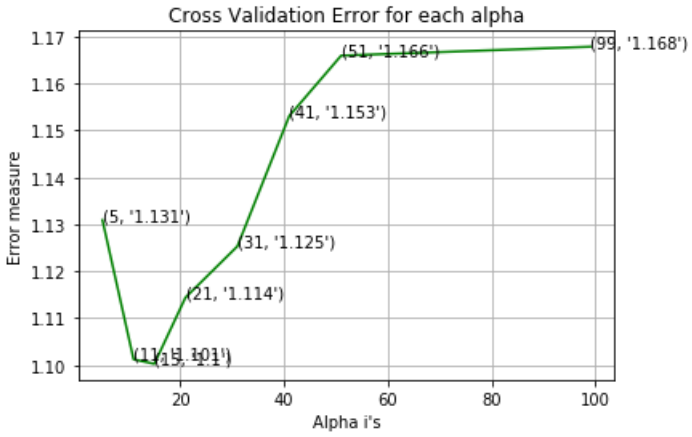
4.2.1. Hyper parameter tuning

In [0]:

```
# 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-ne
ighbors-geometric-intuition-with-a-toy-example-1/
#-----------------------------------


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


alpha = [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)
```

```
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_lo
ss(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, p
redict_y, labels=clf.classes_, eps=1e-15))
```

```
for alpha = 5
Log Loss : 1.1309170126692265
for alpha = 11
Log Loss : 1.1012397762291362
for alpha = 15
Log Loss : 1.1002748803755749
for alpha = 21
Log Loss : 1.1144110925957647
for alpha = 31
Log Loss : 1.1253206995500455
for alpha = 41
Log Loss : 1.1530939773909168
for alpha = 51
Log Loss : 1.1659585643007098
for alpha = 99
Log Loss : 1.1678505034822115
```




```
For values of best alpha =  15 The train log loss is: 0.7056892871225193
For values of best alpha =  15 The cross validation log loss is: 1.1002748803755749
For values of best alpha =  15 The test log loss is: 1.0911901980302394
```

4.2.2. Testing the model with best hyper paramters

In [0]:

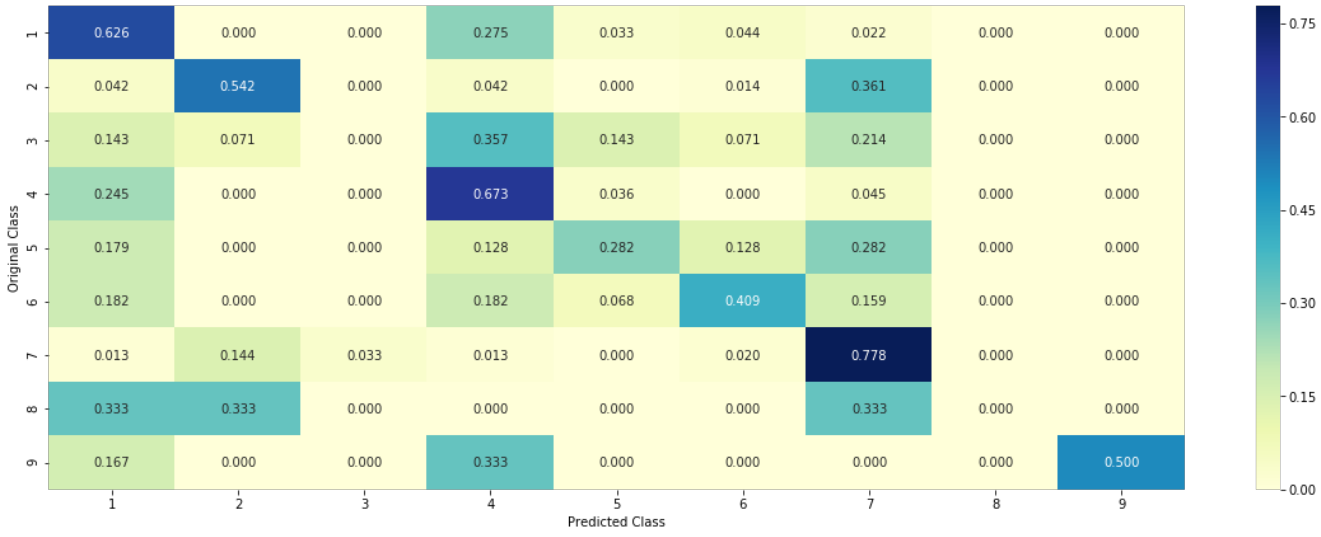
```
# find more about KNeighborsClassifier() here http://scikit-
learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html
# -------------------------
# default parameter
# KNeighborsClassifier(n_neighbors=5, weights='uniform', algorithm='auto', leaf_size=30, p=2,
# metric='minkowski', 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-ne
ighbors-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 : 1.1002748803755749
Number of mis-classified points : 0.3966165413533835
-------------------- Confusion matrix --------------------
```


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


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



4.2.3.Sample Query point -1

In [0]:

```
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].reshape(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 : 6
Actual Class : 7
The  15  nearest neighbours of the test points belongs to classes [7 7 7 6 6 7 6 7 6 7 7 7 7 7 6]
Fequency of nearest points : Counter({7: 10, 6: 5})
```

4.2.4. Sample Query Point-2

In [0]:

```
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].reshape(1, -1), alpha[best_alpha
])
print("the k value for knn is",alpha[best_alpha],"and the nearest neighbours of the test points be
longs to classes",train_y[neighbors[1][0]])
print("Fequency of nearest points :",Counter(train_y[neighbors[1][0]]))
```

```
Predicted Class : 7
Actual Class : 7
the k value for knn is 15 and the nearest neighbours of the test points belongs to classes [2 7 5
7 7 2 7 7 7 2 7 7 7 7 7]
Fequency of nearest points : Counter({7: 11, 2: 3, 5: 1})
```

4.3. Logistic Regression

4.3.1. With Class balancing

4.3.1.1. Hyper paramter tuning

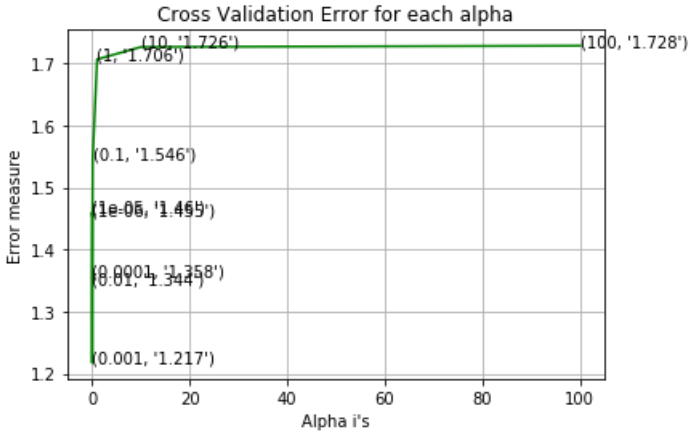
In [0]:

```
# 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_i
ter=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/
#--------------------------------


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

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

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


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

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

```
for alpha = 1e-06
Log Loss : 1.4554353198842396
for alpha = 1e-05
Log Loss : 1.4602144866667575
for alpha = 0.0001
Log Loss : 1.358469527280309
```

```
for alpha = 0.001
Log Loss : 1.217324457704446
for alpha = 0.01
Log Loss : 1.3437838209291793
for alpha = 0.1
Log Loss : 1.5457557924182381
for alpha = 1
Log Loss : 1.706360520395438
for alpha = 10
Log Loss : 1.7261917214695601
for alpha = 100
Log Loss : 1.7282505302427342
```



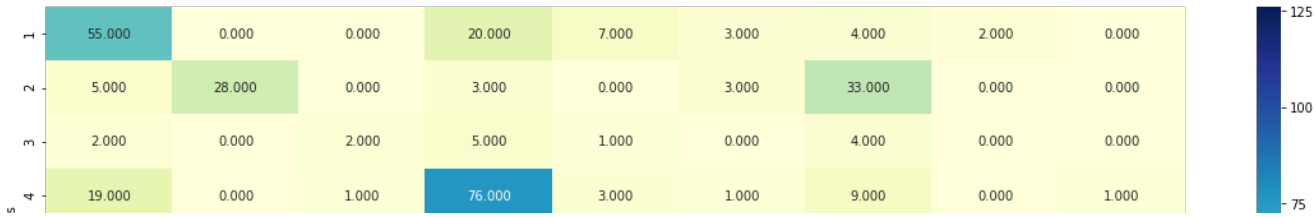

```
For values of best alpha =  0.001 The train log loss is: 0.6153177097029675
For values of best alpha =  0.001 The cross validation log loss is: 1.217324457704446
For values of best alpha =  0.001 The test log loss is: 1.1047257743181136
```

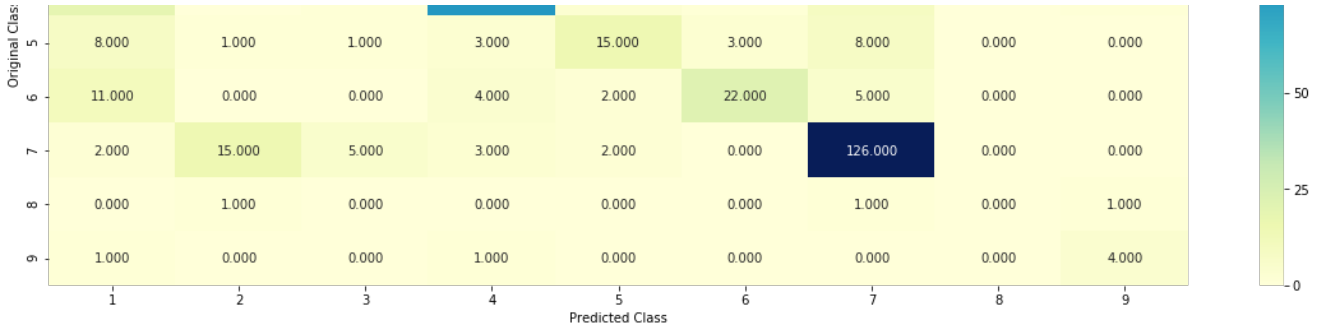
4.3.1.2. Testing the model with best hyper paramters

In [0]:

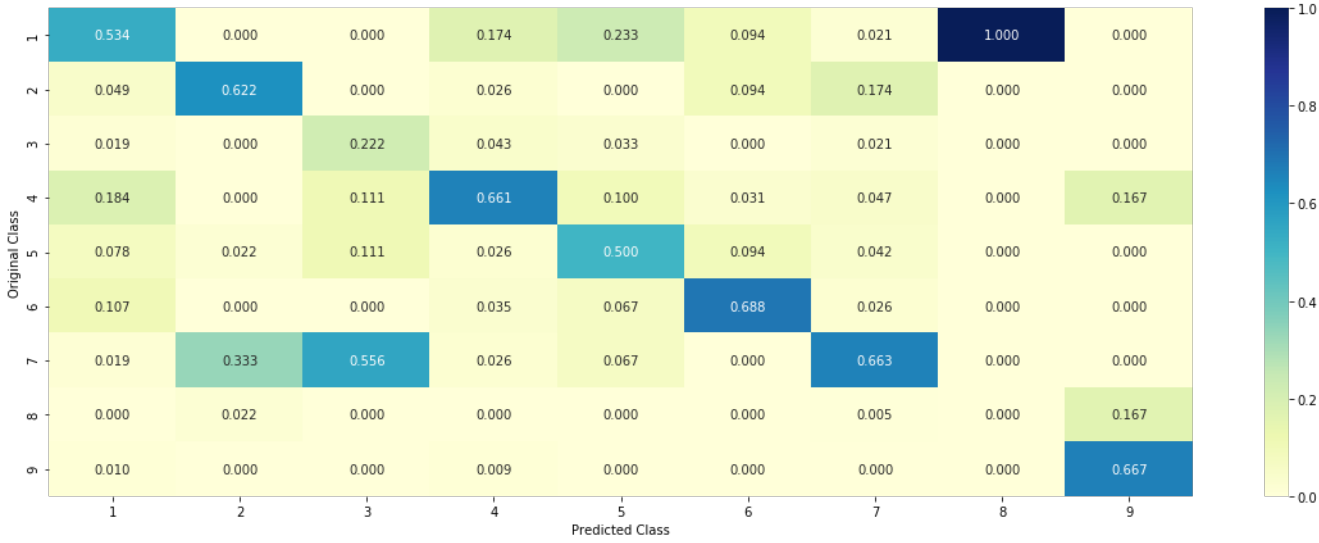
```
# read more about SGDClassifier() at http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDClassifier.html
# ------------------------------
# default parameters
# SGDClassifier(loss='hinge', penalty='l2', alpha=0.0001, l1_ratio=0.15, fit_intercept=True, max_iter=None, tol=None,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

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

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

```
Log loss : 1.217324457704446
Number of mis-classified points : 0.38345864661654133
-------------------- Confusion matrix --------------------
```

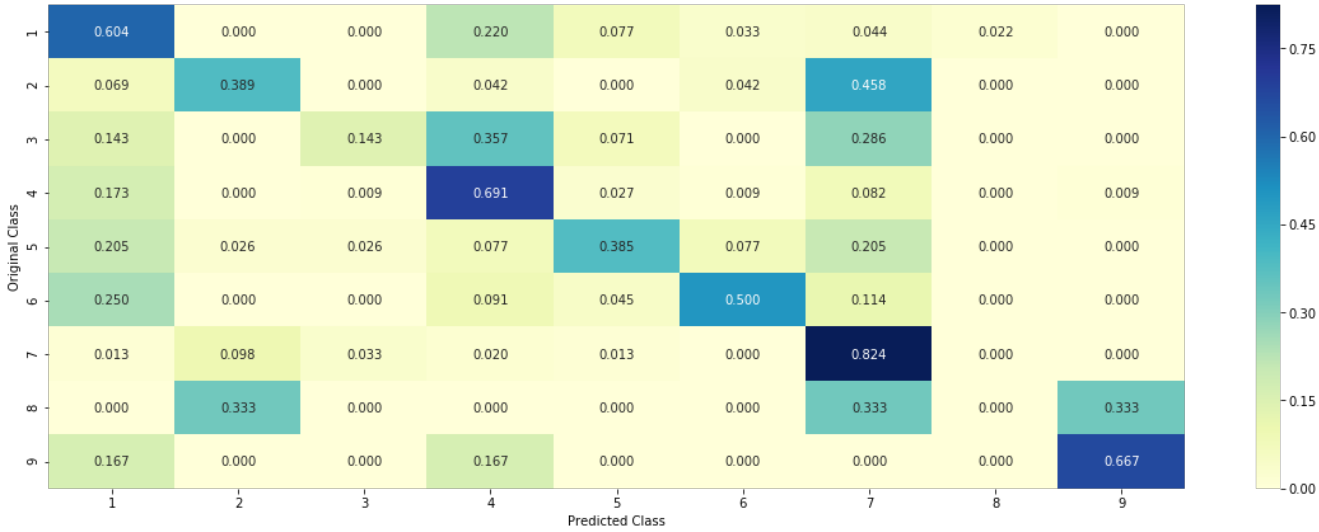




---------- Precision matrix (Columm Sum=1) ----------



---------- 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 query point")
    print("-"*50)
    print("The features that are most importent of the ",predicted_cls[0]," class:")
    print (tabulate(tabulte_list, headers=["Index",'Feature name', 'Present or Not']))
```

4.3.1.3.1. Correctly Classified point

In [0]:

```
# from tabulate import tabulate
clf = SGDClassifier(class_weight='balanced', alpha=alpha[best_alpha], penalty='l2', loss='log', ran
dom_state=42)
clf.fit(train_x_onehotCoding,train_y)
test_point_index = 1
no_feature = 500
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],
test_df['TEXT'].iloc[test_point_index],test_df['Gene'].iloc[test_point_index],test_df['Variation']
.iloc[test_point_index], no_feature)
```

```
Predicted Class : 7
Predicted Class Probabilities: [[0.0045 0.1905 0.0012 0.0012 0.0047 0.0014 0.7872 0.0076 0.0017]]
Actual Class : 7
--------------------------------------------------
23 Text feature [constitutively] present in test data point [True]
39 Text feature [flt1] present in test data point [True]
79 Text feature [oncogene] present in test data point [True]
80 Text feature [oncogenes] present in test data point [True]
84 Text feature [cysteine] present in test data point [True]
89 Text feature [inhibited] present in test data point [True]
137 Text feature [technology] present in test data point [True]
160 Text feature [dramatic] present in test data point [True]
162 Text feature [gaiix] present in test data point [True]
166 Text feature [ligand] present in test data point [True]
177 Text feature [downstream] present in test data point [True]
181 Text feature [concentrations] present in test data point [True]
182 Text feature [thyroid] present in test data point [True]
187 Text feature [expressing] present in test data point [True]
217 Text feature [activating] present in test data point [True]
241 Text feature [cdnas] present in test data point [True]
250 Text feature [manageable] present in test data point [True]
265 Text feature [axilla] present in test data point [True]
302 Text feature [inhibitor] present in test data point [True]
311 Text feature [cot] present in test data point [True]
313 Text feature [viability] present in test data point [True]
334 Text feature [activation] present in test data point [True]
352 Text feature [forced] present in test data point [True]
368 Text feature [subcutaneous] present in test data point [True]
371 Text feature [melanocyte] present in test data point [True]
376 Text feature [erk1] present in test data point [True]
388 Text feature [hours] present in test data point [True]
446 Text feature [procure] present in test data point [True]
448 Text feature [doses] present in test data point [True]
480 Text feature [mapk] present in test data point [True]
Out of the top  500  features  30 are present in query point
```

4.3.1.3.2. Incorrectly Classified point

In [0]:

```
test_point_index = 100
no_feature = 500
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],
test_df['TEXT'].iloc[test_point_index],test_df['Gene'].iloc[test_point_index],test_df['Variation']
.iloc[test_point_index], no_feature)
```

```
Predicted Class : 7
Predicted Class Probabilities: [[0.0482 0.2032 0.0108 0.0446 0.071  0.0164 0.5932 0.0078 0.0046]]
Actual Class : 7
--------------------------------------------------
23 Text feature [constitutively] present in test data point [True]
29 Text feature [constitutive] present in test data point [True]
47 Text feature [activated] present in test data point [True]
79 Text feature [oncogene] present in test data point [True]
89 Text feature [inhibited] present in test data point [True]
93 Text feature [transforming] present in test data point [True]
108 Text feature [transform] present in test data point [True]
148 Text feature [receptors] present in test data point [True]
177 Text feature [downstream] present in test data point [True]
210 Text feature [isozyme] present in test data point [True]
217 Text feature [activating] present in test data point [True]
232 Text feature [exchange] present in test data point [True]
326 Text feature [murine] present in test data point [True]
333 Text feature [agar] present in test data point [True]
334 Text feature [activation] present in test data point [True]
Out of the top  500  features  15 are present in query point
```

4.3.2. Without Class balancing

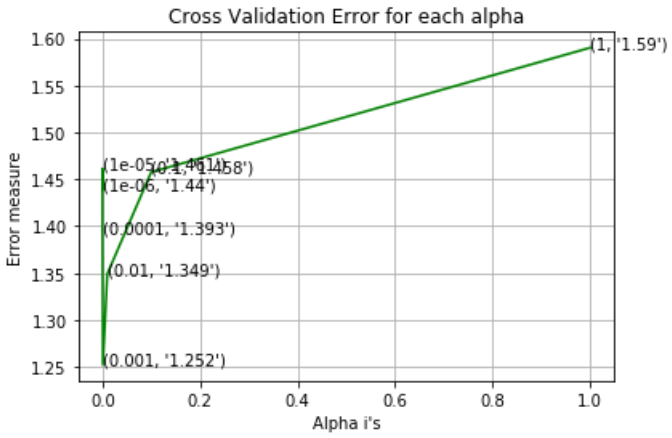
4.3.2.1. Hyper paramter tuning

In [0]:

```
# 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_i
ter=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-in
tuition-1/
#-------------------------------



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

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

alpha = [10 ** x for x in range(-6, 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_lo
ss(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, p
redict_y, labels=clf.classes_, eps=1e-15))
```

```
for alpha = 1e-06
Log Loss : 1.4395190222240433
for alpha = 1e-05
Log Loss : 1.4613951945118617
for alpha = 0.0001
Log Loss : 1.392640595913179
for alpha = 0.001
Log Loss : 1.2521811628755943
for alpha = 0.01
Log Loss : 1.349151219922669
for alpha = 0.1
Log Loss : 1.457591708320943
for alpha = 1
Log Loss : 1.5902258764770603
```



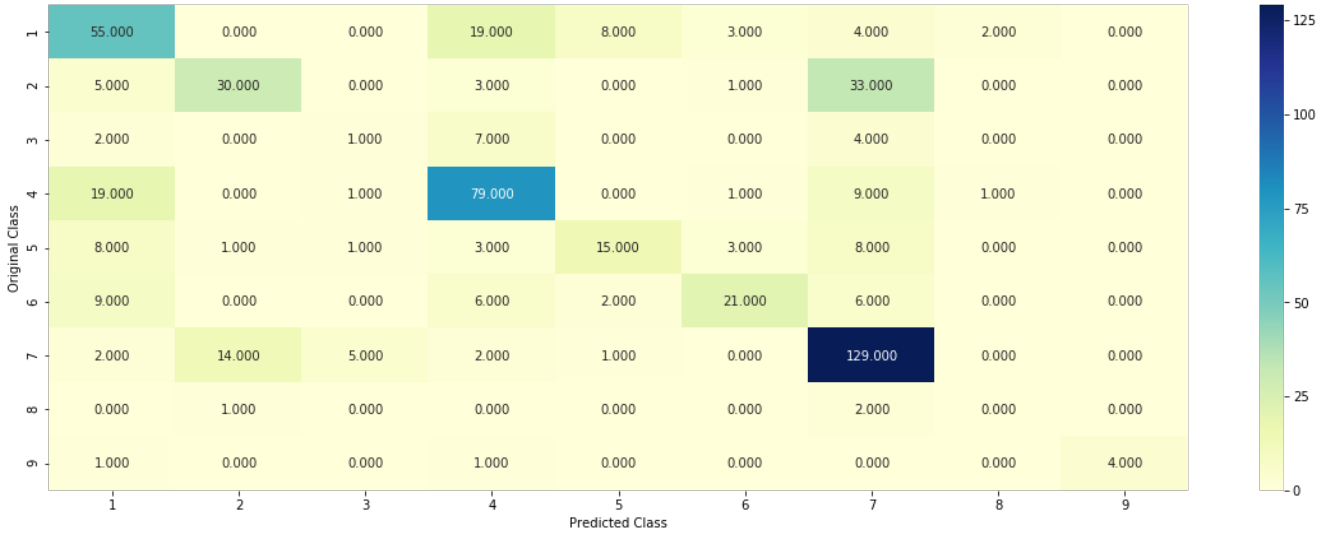
```
For values of best alpha =  0.001 The train log loss is: 0.6257422677412771
For values of best alpha =  0.001 The cross validation log loss is: 1.2521811628755943
For values of best alpha =  0.001 The test log loss is: 1.1306020069615057
```

4.3.2.2. Testing model with best hyper parameters

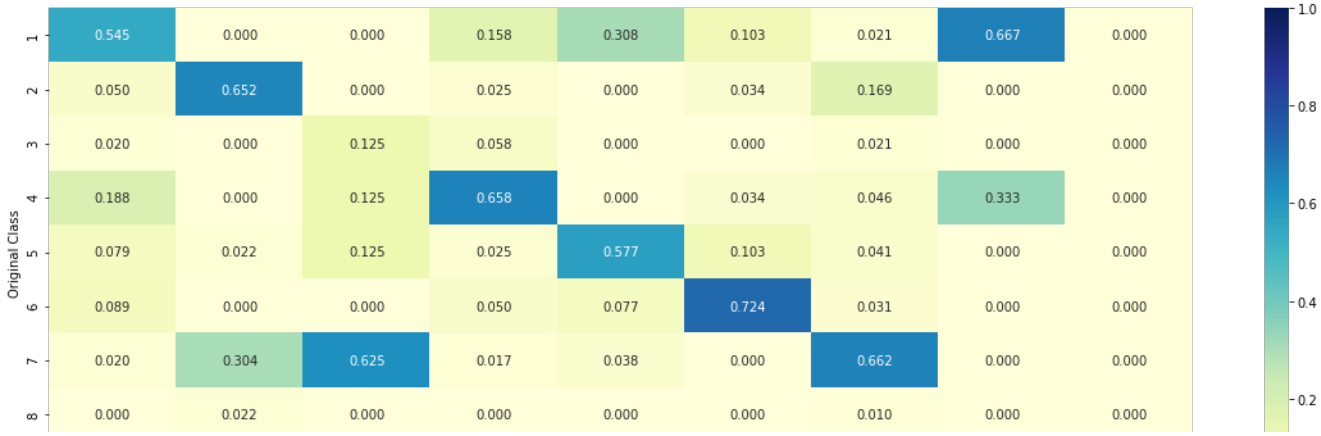
In [0]:

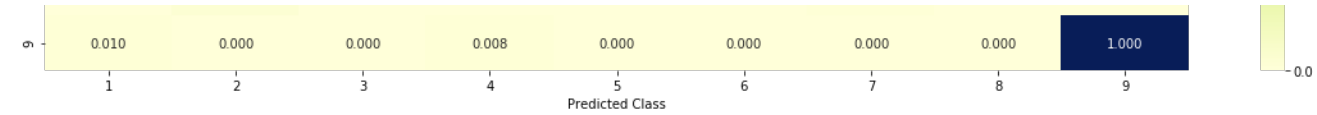
```
# read more about SGDClassifier() at http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDClassifier.html
# ------------------------------
# default parameters
# SGDClassifier(loss='hinge', penalty='l2', alpha=0.0001, l1_ratio=0.15, fit_intercept=True, max_iter=None, tol=None,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

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

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

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

```
Log loss : 1.2521811628755943
Number of mis-classified points : 0.37218045112781956
-------------------- Confusion matrix --------------------
```

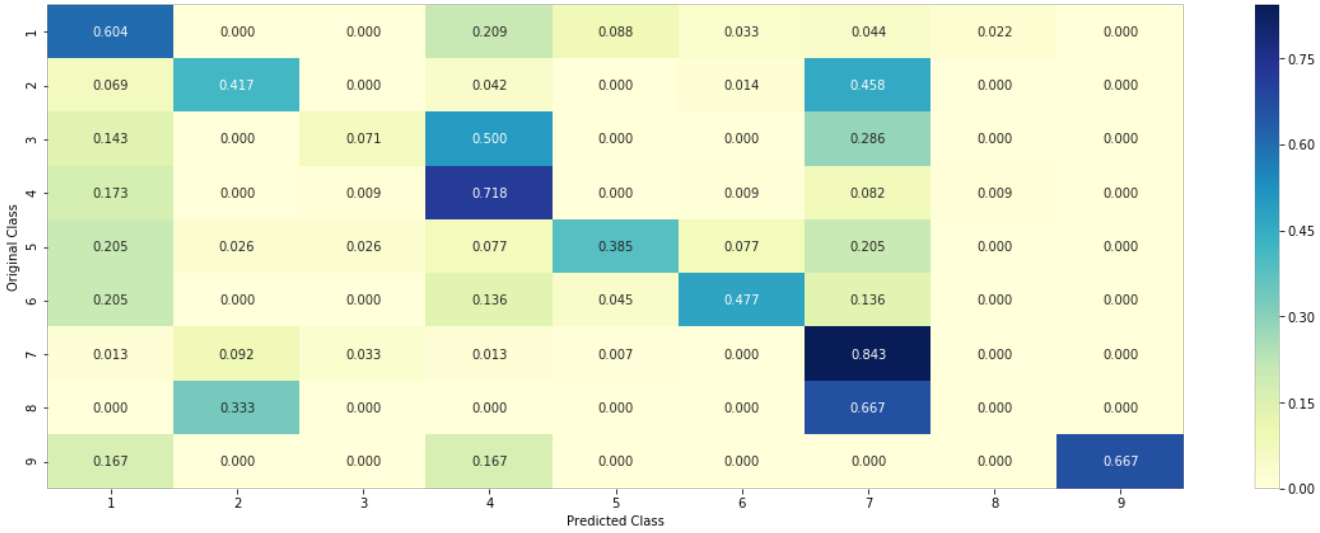



```
-------------------- Precision matrix (Columm Sum=1) --------------------
```






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



4.3.2.3. Feature Importance, Correctly Classified point

In [0]:

```
clf = SGDClassifier(alpha=alpha[best_alpha], penalty='l2', loss='log', random_state=42)
clf.fit(train_x_onehotCoding,train_y)
test_point_index = 1
no_feature = 500
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],
test_df['TEXT'].iloc[test_point_index],test_df['Gene'].iloc[test_point_index],test_df['Variation']
.iloc[test_point_index], no_feature)
```

```
Predicted Class : 7
Predicted Class Probabilities: [[5.100e-03 1.255e-01 2.000e-04 1.300e-03 2.300e-03 1.400e-03 8.556
e-01
  8.500e-03 1.000e-04]]
Actual Class : 7
--------------------------------------------------
60 Text feature [constitutively] present in test data point [True]
107 Text feature [flt1] present in test data point [True]
124 Text feature [cysteine] present in test data point [True]
157 Text feature [oncogenes] present in test data point [True]
158 Text feature [inhibited] present in test data point [True]
195 Text feature [activating] present in test data point [True]
200 Text feature [ligand] present in test data point [True]
203 Text feature [oncogene] present in test data point [True]
204 Text feature [technology] present in test data point [True]
257 Text feature [gaiix] present in test data point [True]
260 Text feature [concentrations] present in test data point [True]
265 Text feature [downstream] present in test data point [True]
314 Text feature [hki] present in test data point [True]
316 Text feature [dramatic] present in test data point [True]
323 Text feature [expressing] present in test data point [True]
371 Text feature [cdnas] present in test data point [True]
380 Text feature [viability] present in test data point [True]
412 Text feature [thyroid] present in test data point [True]
459 Text feature [activation] present in test data point [True]
461 Text feature [manageable] present in test data point [True]
```

```
462 Text feature [ser473] present in test data point [True]
468 Text feature [axilla] present in test data point [True]
495 Text feature [extracellular] present in test data point [True]
Out of the top  500  features  23 are present in query point
```

4.3.2.4. Feature Importance, Inorrectly Classified point

In [0]:

```
test_point_index = 100
no_feature = 500
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],
test_df['TEXT'].iloc[test_point_index],test_df['Gene'].iloc[test_point_index],test_df['Variation']
.iloc[test_point_index], no_feature)
```

```
Predicted Class : 7
Predicted Class Probabilities: [[0.0485 0.1851 0.0052 0.0442 0.0617 0.0143 0.6317 0.0072 0.0022]]
Actual Class : 7
--------------------------------------------------
60 Text feature [constitutively] present in test data point [True]
89 Text feature [constitutive] present in test data point [True]
116 Text feature [activated] present in test data point [True]
158 Text feature [inhibited] present in test data point [True]
159 Text feature [transforming] present in test data point [True]
193 Text feature [receptors] present in test data point [True]
195 Text feature [activating] present in test data point [True]
203 Text feature [oncogene] present in test data point [True]
226 Text feature [transform] present in test data point [True]
241 Text feature [isozyme] present in test data point [True]
265 Text feature [downstream] present in test data point [True]
377 Text feature [agar] present in test data point [True]
442 Text feature [interatomic] present in test data point [True]
459 Text feature [activation] present in test data point [True]
Out of the top  500  features  14 are present in query point
```

4.4. Linear Support Vector Machines

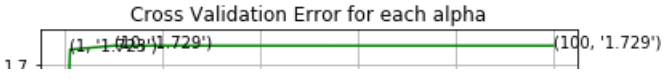
4.4.1. Hyper paramter tuning

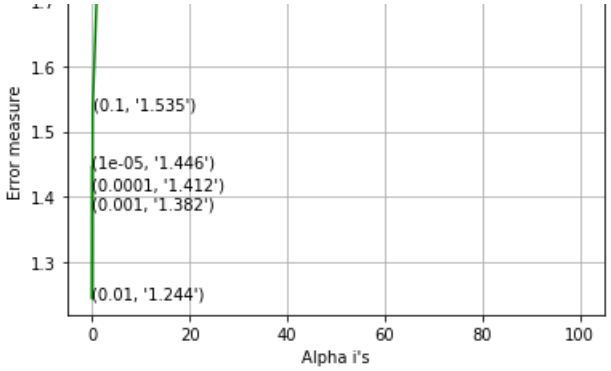
In [0]:

```
# 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, t
ol=0.001,
# cache_size=200, class_weight=None, verbose=False, max_iter=-1, decision_function_shape='ovr', ra
ndom_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, method='sigmoid', cv=3)
#
# some of the methods of CalibratedClassifierCV()
# fit(X, y[, sample_weight]) Fit the calibrated model
# get_params([deep]) Get parameters for this estimator.
# predict(X) Predict the target of new samples.
# predict_proba(X) Posterior probabilities of classification
#-------------------------------------
# video link:
#-------------------------------------

alpha = [10 ** x for x in range(-5, 3)]
cv_log_error_array = []
for i in alpha:
    print("for C =", i)
#     clf = SVC(C=i,kernel='linear',probability=True, class_weight='balanced')
    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()


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', r
andom_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_lo
ss(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, p
redict_y, labels=clf.classes_, eps=1e-15))
```

```
for C = 1e-05
Log Loss : 1.4456349250609233
for C = 0.0001
Log Loss : 1.4117883301099556
for C = 0.001
Log Loss : 1.3818342037841624
for C = 0.01
Log Loss : 1.2442964974823838
for C = 0.1
Log Loss : 1.5346828298587332
for C = 1
Log Loss : 1.722800653929441
for C = 10
Log Loss : 1.7286360420759161
for C = 100
Log Loss : 1.7286184454094997
```






```
For values of best alpha =  0.01 The train log loss is: 0.7628309867716067
For values of best alpha =  0.01 The cross validation log loss is: 1.2442964974823838
For values of best alpha =  0.01 The test log loss is: 1.1541891969863685
```

4.4.2. Testing model with best hyper parameters

In [0]:

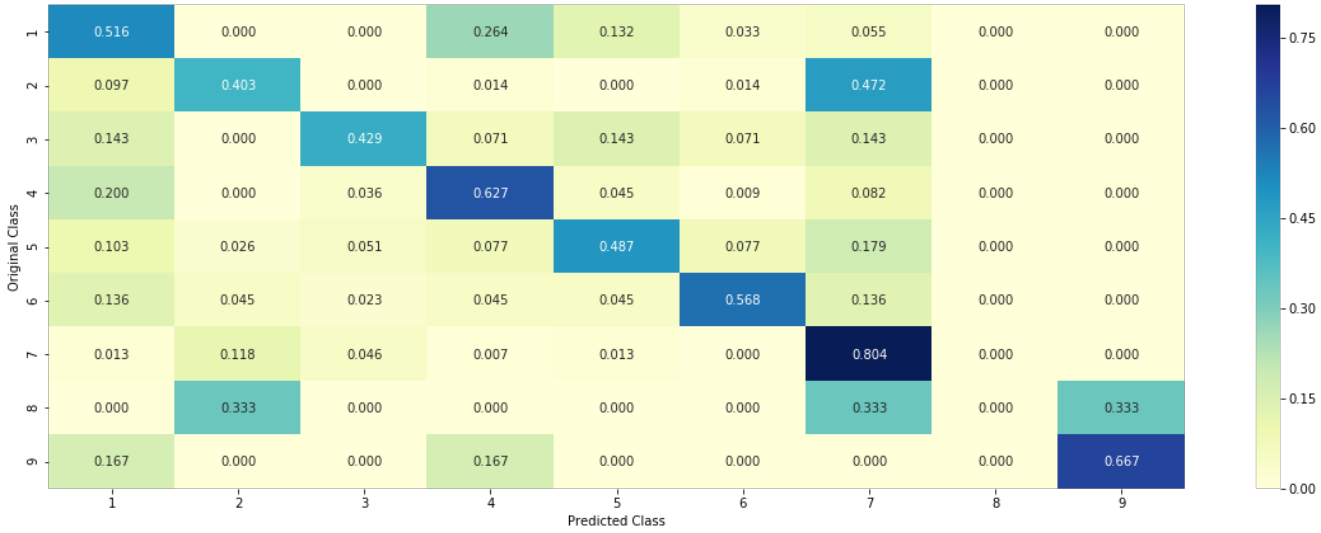
```
# read more about support vector machines with linear kernals here http://scikit-
learn.org/stable/modules/generated/sklearn.svm.SVC.html

# --------------------------------
# default parameters
# SVC(C=1.0, kernel='rbf', degree=3, gamma='auto', coef0=0.0, shrinking=True, probability=False, t
ol=0.001,
# cache_size=200, class_weight=None, verbose=False, max_iter=-1, decision_function_shape='ovr', ra
ndom_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/
# --------------------------------


# 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_onehotCoding,cv_y, clf)
```

```
Log loss : 1.2442964974823838
Number of mis-classified points : 0.39473684210526316
-------------------- Confusion matrix --------------------
```

Precision matrix (Column Sum=1)

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

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



4.3.3. Feature Importance

4.3.3.1. For Correctly classified point

In [0]:

```
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 = 500
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],
test_df['TEXT'].iloc[test_point_index],test_df['Gene'].iloc[test_point_index],test_df['Variation']
.iloc[test_point_index], no_feature)
```

```
Predicted Class : 7
Predicted Class Probabilities: [[0.0153 0.1199 0.0029 0.0151 0.0121 0.0075 0.8104 0.0129 0.0039]]
Actual Class : 7
--------------------------------------------------
```

```
--------------------------------------------------
28 Text feature [constitutively] present in test data point [True]
29 Text feature [cysteine] present in test data point [True]
49 Text feature [cdnas] present in test data point [True]
76 Text feature [flt1] present in test data point [True]
79 Text feature [concentrations] present in test data point [True]
82 Text feature [gaiix] present in test data point [True]
96 Text feature [technology] present in test data point [True]
101 Text feature [inhibited] present in test data point [True]
104 Text feature [activating] present in test data point [True]
114 Text feature [oncogenes] present in test data point [True]
147 Text feature [expressing] present in test data point [True]
150 Text feature [mapk] present in test data point [True]
151 Text feature [oncogene] present in test data point [True]
169 Text feature [thyroid] present in test data point [True]
171 Text feature [inhibitor] present in test data point [True]
205 Text feature [transduced] present in test data point [True]
211 Text feature [seeded] present in test data point [True]
230 Text feature [ligand] present in test data point [True]
255 Text feature [activation] present in test data point [True]
279 Text feature [downstream] present in test data point [True]
314 Text feature [doses] present in test data point [True]
351 Text feature [subcutaneous] present in test data point [True]
366 Text feature [atcc] present in test data point [True]
405 Text feature [melanocyte] present in test data point [True]
436 Text feature [hours] present in test data point [True]
445 Text feature [selleck] present in test data point [True]
446 Text feature [dramatic] present in test data point [True]
454 Text feature [chemiluminescence] present in test data point [True]
487 Text feature [viability] present in test data point [True]
489 Text feature [ser473] present in test data point [True]
Out of the top  500  features  30 are present in query point
```

4.3.3.2. For Incorrectly classified point

In [0]:

```
test_point_index = 100
no_feature = 500
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],
test_df['TEXT'].iloc[test_point_index],test_df['Gene'].iloc[test_point_index],test_df['Variation']
.iloc[test_point_index], no_feature)
```

```
Predicted Class : 7
Predicted Class Probabilities: [[0.0786 0.1516 0.0146 0.1064 0.1105 0.0323 0.4839 0.0128 0.0094]]
Actual Class : 7
--------------------------------------------------
28 Text feature [constitutively] present in test data point [True]
40 Text feature [constitutive] present in test data point [True]
73 Text feature [activated] present in test data point [True]
75 Text feature [transforming] present in test data point [True]
94 Text feature [receptors] present in test data point [True]
97 Text feature [exchange] present in test data point [True]
101 Text feature [inhibited] present in test data point [True]
104 Text feature [activating] present in test data point [True]
151 Text feature [oncogene] present in test data point [True]
231 Text feature [transform] present in test data point [True]
255 Text feature [activation] present in test data point [True]
279 Text feature [downstream] present in test data point [True]
440 Text feature [doubled] present in test data point [True]
470 Text feature [substituting] present in test data point [True]
Out of the top  500  features  14 are present in query point
```

4.5 Random Forest Classifier

4.5.1. Hyper paramter tuning (With One hot Encoding)

In [0]:

```
# --------------------------------
# default parameters
# sklearn.ensemble.RandomForestClassifier(n_estimators=10, criterion='gini', max_depth=None, min_s
amples_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-fores
t-and-their-construction-2/
# --------------------------------


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

alpha = [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]).ravel()
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)),
(features[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)], criterion='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))
```

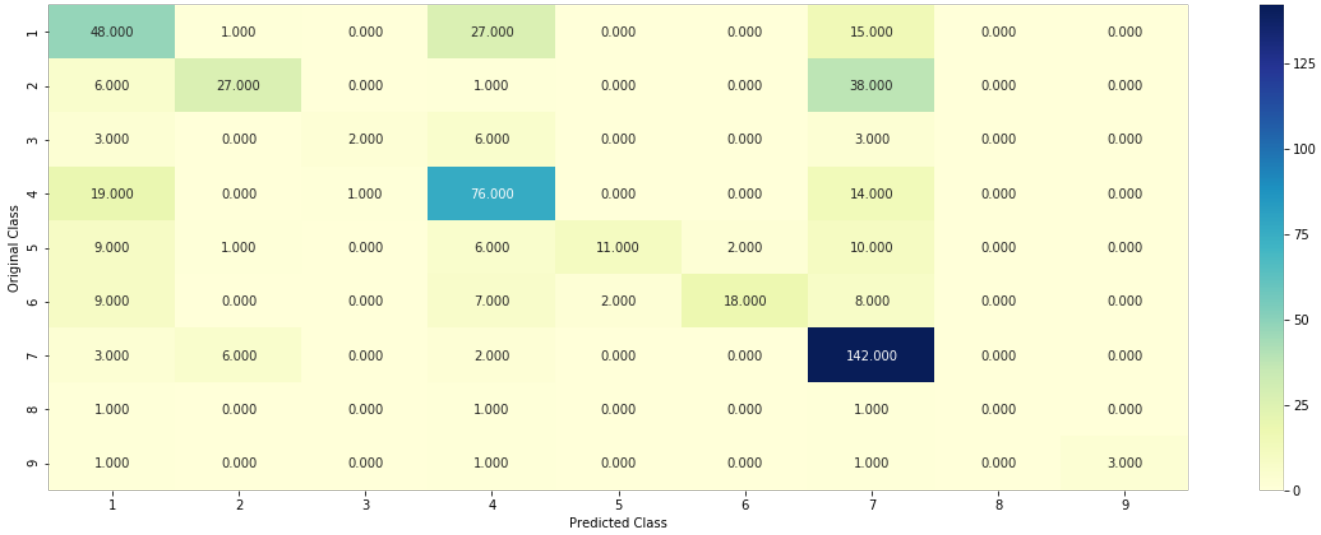
```
for n_estimators = 100 and max depth =  5
Log Loss : 1.2572535683354957
for n_estimators = 100 and max depth =  10
Log Loss : 1.1868414223711878
for n_estimators = 200 and max depth =  5
Log Loss : 1.2378734502517341
for n_estimators = 200 and max depth =  10
Log Loss : 1.1811031780258958
for n_estimators = 500 and max depth =  5
Log Loss : 1.2368241894319212
for n_estimators = 500 and max depth =  10
Log Loss : 1.176754594516683
for n_estimators = 1000 and max depth =  5
Log Loss : 1.2357829533963691
for n_estimators = 1000 and max depth =  10
Log Loss : 1.174993079576866
for n_estimators = 2000 and max depth =  5
Log Loss : 1.236042392554891
for n_estimators = 2000 and max depth =  10
Log Loss : 1.1759745074379755
For values of best estimator =  1000 The train log loss is: 0.7095396732082752
For values of best estimator =  1000 The cross validation log loss is: 1.174993079576866
For values of best estimator =  1000 The test log loss is: 1.1630923149103904
```

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

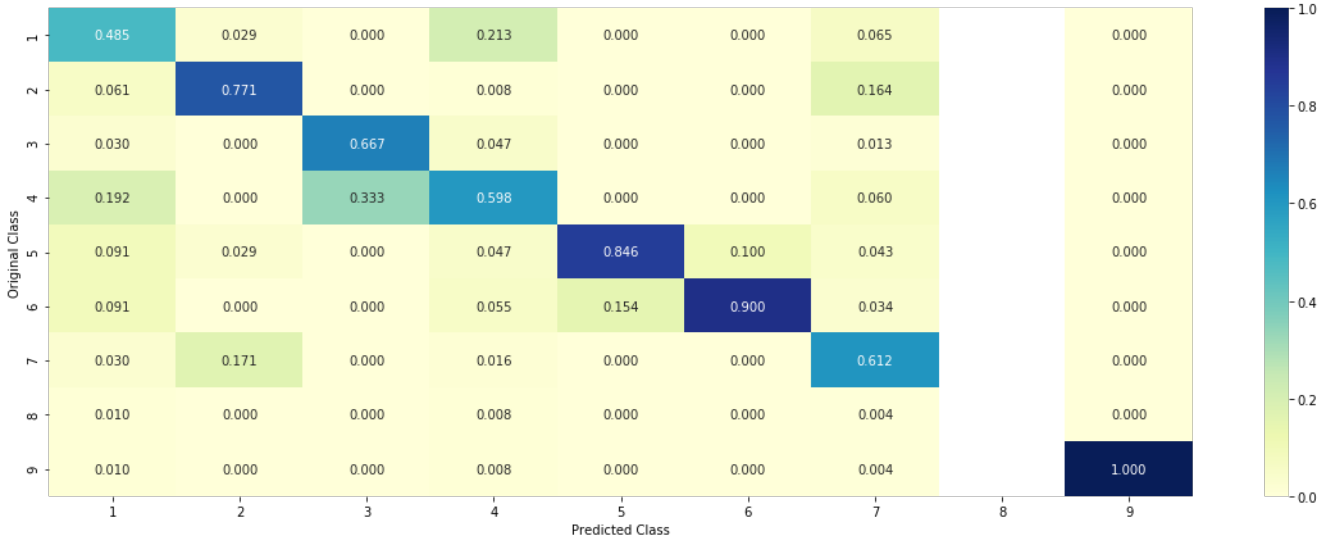
In [0]:

```
# --------------------------------
# default parameters
# sklearn.ensemble.RandomForestClassifier(n_estimators=10, criterion='gini', max_depth=None, min_s
amples_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-fores
t-and-their-construction-2/
# --------------------------------

clf = RandomForestClassifier(n_estimators=alpha[int(best_alpha/2)], criterion='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_onehotCoding,cv_y, clf)
```

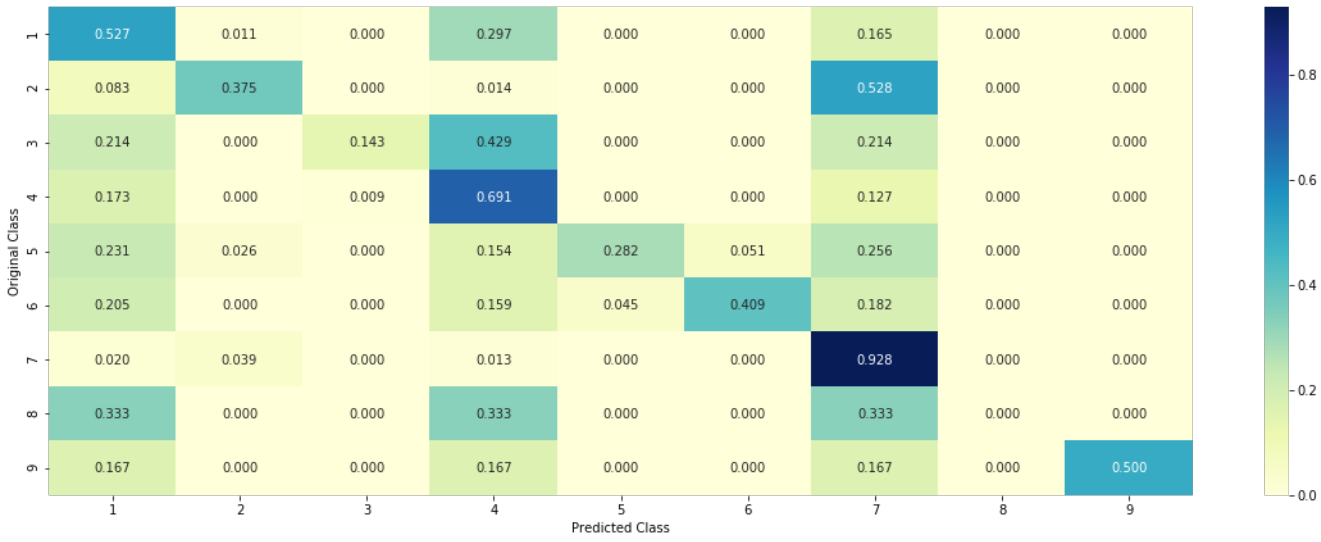
```
Log loss : 1.174993079576866
Number of mis-classified points : 0.38533834586466165
-------------------- Confusion matrix --------------------
```



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



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



4.5.3. Feature Importance

4.5.3.1. Correctly Classified point

In [0]:

```
# test_point_index = 10
clf = RandomForestClassifier(n_estimators=alpha[int(best_alpha/2)], criterion='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 = 100
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], test_df['TEXT'].iloc[test_point_index],test_df['Gene'].
iloc[test_point_index],test_df['Variation'].iloc[test_point_index], no_feature)
```

```
Predicted Class : 7
Predicted Class Probabilities: [[0.0454 0.1404 0.0133 0.029  0.036  0.0294 0.6977 0.005  0.004 ]]
Actual Class : 7
--------------------------------------------------
0 Text feature [inhibitors] present in test data point [True]
1 Text feature [kinase] present in test data point [True]
2 Text feature [activating] present in test data point [True]
3 Text feature [tyrosine] present in test data point [True]
4 Text feature [missense] present in test data point [True]
5 Text feature [inhibitor] present in test data point [True]
7 Text feature [treatment] present in test data point [True]
8 Text feature [oncogenic] present in test data point [True]
9 Text feature [suppressor] present in test data point [True]
10 Text feature [activation] present in test data point [True]
11 Text feature [phosphorylation] present in test data point [True]
12 Text feature [kinases] present in test data point [True]
13 Text feature [nonsense] present in test data point [True]
14 Text feature [akt] present in test data point [True]
15 Text feature [function] present in test data point [True]
17 Text feature [erk] present in test data point [True]
19 Text feature [growth] present in test data point [True]
20 Text feature [variants] present in test data point [True]
22 Text feature [frameshift] present in test data point [True]
24 Text feature [therapeutic] present in test data point [True]
25 Text feature [functional] present in test data point [True]
28 Text feature [signaling] present in test data point [True]
30 Text feature [patients] present in test data point [True]
31 Text feature [cells] present in test data point [True]
32 Text feature [constitutively] present in test data point [True]
34 Text feature [trials] present in test data point [True]
35 Text feature [therapy] present in test data point [True]
37 Text feature [erk1] present in test data point [True]
38 Text feature [activate] present in test data point [True]
39 Text feature [downstream] present in test data point [True]
41 Text feature [efficacy] present in test data point [True]
42 Text feature [protein] present in test data point [True]
43 Text feature [loss] present in test data point [True]
44 Text feature [inhibited] present in test data point [True]
45 Text feature [expressing] present in test data point [True]
46 Text feature [pten] present in test data point [True]
48 Text feature [lines] present in test data point [True]
49 Text feature [treated] present in test data point [True]
50 Text feature [proliferation] present in test data point [True]
51 Text feature [drug] present in test data point [True]
57 Text feature [mek] present in test data point [True]
59 Text feature [inhibition] present in test data point [True]
61 Text feature [repair] present in test data point [True]
62 Text feature [sensitivity] present in test data point [True]
64 Text feature [receptor] present in test data point [True]
66 Text feature [assays] present in test data point [True]
68 Text feature [survival] present in test data point [True]
69 Text feature [cell] present in test data point [True]
71 Text feature [ligand] present in test data point [True]
73 Text feature [expression] present in test data point [True]
```

```
74 Text feature [variant] present in test data point [True]
75 Text feature [oncogene] present in test data point [True]
78 Text feature [extracellular] present in test data point [True]
79 Text feature [doses] present in test data point [True]
80 Text feature [mapk] present in test data point [True]
81 Text feature [hours] present in test data point [True]
84 Text feature [information] present in test data point [True]
86 Text feature [harboring] present in test data point [True]
90 Text feature [dna] present in test data point [True]
91 Text feature [concentrations] present in test data point [True]
92 Text feature [likelihood] present in test data point [True]
93 Text feature [months] present in test data point [True]
94 Text feature [binding] present in test data point [True]
96 Text feature [imatinib] present in test data point [True]
98 Text feature [preclinical] present in test data point [True]
Out of the top  100  features  65 are present in query point
```

4.5.3.2. Inorrectly Classified point

In [0]:

```
test_point_index = 100
no_feature = 100
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], test_df['TEXT'].iloc[test_point_index],test_df['Gene'].
iloc[test_point_index],test_df['Variation'].iloc[test_point_index], no_feature)
```

```
Predicted Class : 7
Predicted Class Probabilities: [[0.1337 0.116  0.0224 0.1773 0.0674 0.0545 0.4156 0.0071 0.0059]]
Actuall Class : 7
--------------------------------------------------
0 Text feature [inhibitors] present in test data point [True]
1 Text feature [kinase] present in test data point [True]
2 Text feature [activating] present in test data point [True]
3 Text feature [tyrosine] present in test data point [True]
6 Text feature [activated] present in test data point [True]
8 Text feature [oncogenic] present in test data point [True]
10 Text feature [activation] present in test data point [True]
11 Text feature [phosphorylation] present in test data point [True]
12 Text feature [kinases] present in test data point [True]
14 Text feature [akt] present in test data point [True]
15 Text feature [function] present in test data point [True]
19 Text feature [growth] present in test data point [True]
21 Text feature [constitutive] present in test data point [True]
25 Text feature [functional] present in test data point [True]
28 Text feature [signaling] present in test data point [True]
31 Text feature [cells] present in test data point [True]
32 Text feature [constitutively] present in test data point [True]
38 Text feature [activate] present in test data point [True]
39 Text feature [downstream] present in test data point [True]
42 Text feature [protein] present in test data point [True]
43 Text feature [loss] present in test data point [True]
44 Text feature [inhibited] present in test data point [True]
46 Text feature [pten] present in test data point [True]
47 Text feature [transforming] present in test data point [True]
48 Text feature [lines] present in test data point [True]
50 Text feature [proliferation] present in test data point [True]
53 Text feature [neutral] present in test data point [True]
55 Text feature [transform] present in test data point [True]
56 Text feature [stability] present in test data point [True]
58 Text feature [transformation] present in test data point [True]
59 Text feature [inhibition] present in test data point [True]
62 Text feature [sensitivity] present in test data point [True]
64 Text feature [receptor] present in test data point [True]
66 Text feature [assays] present in test data point [True]
69 Text feature [cell] present in test data point [True]
75 Text feature [oncogene] present in test data point [True]
84 Text feature [information] present in test data point [True]
90 Text feature [dna] present in test data point [True]
```

```
90 Text feature [dna] present in test data point [True]
94 Text feature [binding] present in test data point [True]
Out of the top  100  features  39 are present in query point
```

4.5.3. Hyper paramter tuning (With Response Coding)

In [0]:

```
# --------------------------------
# default parameters
# sklearn.ensemble.RandomForestClassifier(n_estimators=10, criterion='gini', max_depth=None, min_s
amples_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-fores
t-and-their-construction-2/
# --------------------------------


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

alpha = [10,50,100,200,500,1000]
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]).ravel()
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)),
(features[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)], criterion='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 : 2.2657048897349608
for n_estimators = 10 and max depth =  3
Log Loss : 1.7459205010556096
for n_estimators = 10 and max depth =  5
Log Loss : 1.4368353925512503
for n_estimators = 10 and max depth =  10
Log Loss : 1.904597809032912
for n_estimators = 50 and max depth =  2
Log Loss : 1.7221951095007484
for n_estimators = 50 and max depth =  3
Log Loss : 1.4984825877845531
for n_estimators = 50 and max depth =  5
Log Loss : 1.4593628982873716
for n_estimators = 50 and max depth =  10
Log Loss : 1.8434939703555409
for n_estimators = 100 and max depth =  2
Log Loss : 1.6182209245331227
for n_estimators = 100 and max depth =  3
Log Loss : 1.5199297988828253
for n_estimators = 100 and max depth =  5
Log Loss : 1.4177501184246677
for n_estimators = 100 and max depth =  10
Log Loss : 1.8227504417195126
for n_estimators = 200 and max depth =  2
Log Loss : 1.6622571648074496
for n_estimators = 200 and max depth =  3
Log Loss : 1.4800771339141767
for n_estimators = 200 and max depth =  5
Log Loss : 1.4412060242341358
for n_estimators = 200 and max depth =  10
Log Loss : 1.7892406351442258
for n_estimators = 500 and max depth =  2
Log Loss : 1.715950314170445
for n_estimators = 500 and max depth =  3
Log Loss : 1.5658682738699774
for n_estimators = 500 and max depth =  5
Log Loss : 1.4445360301518217
for n_estimators = 500 and max depth =  10
Log Loss : 1.8421097596928397
for n_estimators = 1000 and max depth =  2
Log Loss : 1.6834927870864949
for n_estimators = 1000 and max depth =  3
Log Loss : 1.5631973035931377
for n_estimators = 1000 and max depth =  5
Log Loss : 1.4449980792724129
for n_estimators = 1000 and max depth =  10
Log Loss : 1.85233132619749
For values of best alpha =  100 The train log loss is: 0.060702709444608406
For values of best alpha =  100 The cross validation log loss is: 1.417750118424668
For values of best alpha =  100 The test log loss is: 1.3806278998341923
```

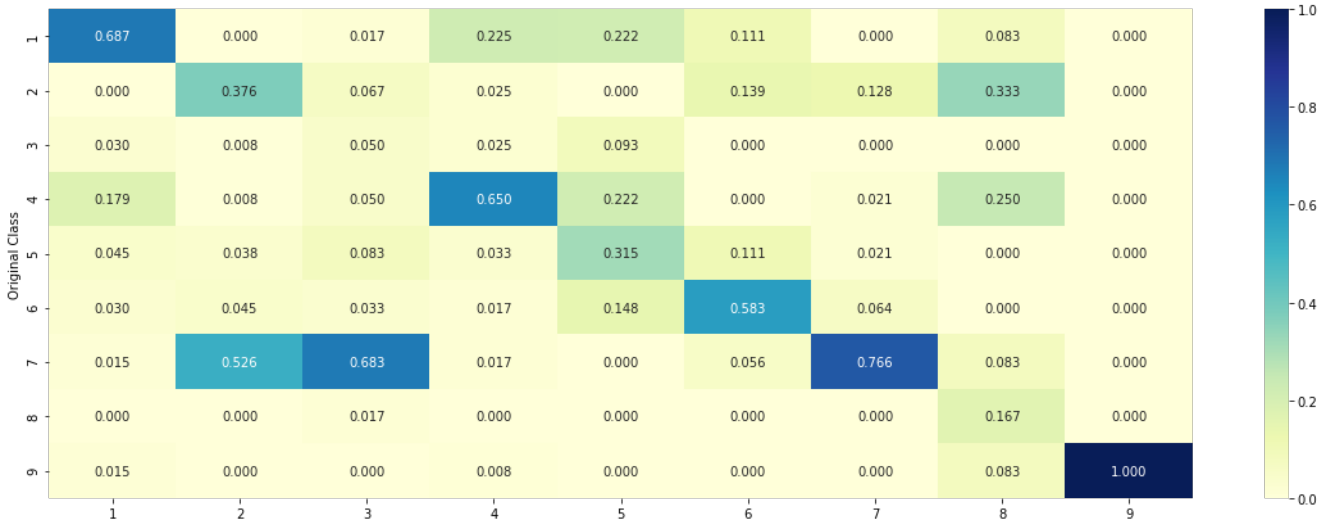
4.5.4. Testing model with best hyper parameters (Response Coding)

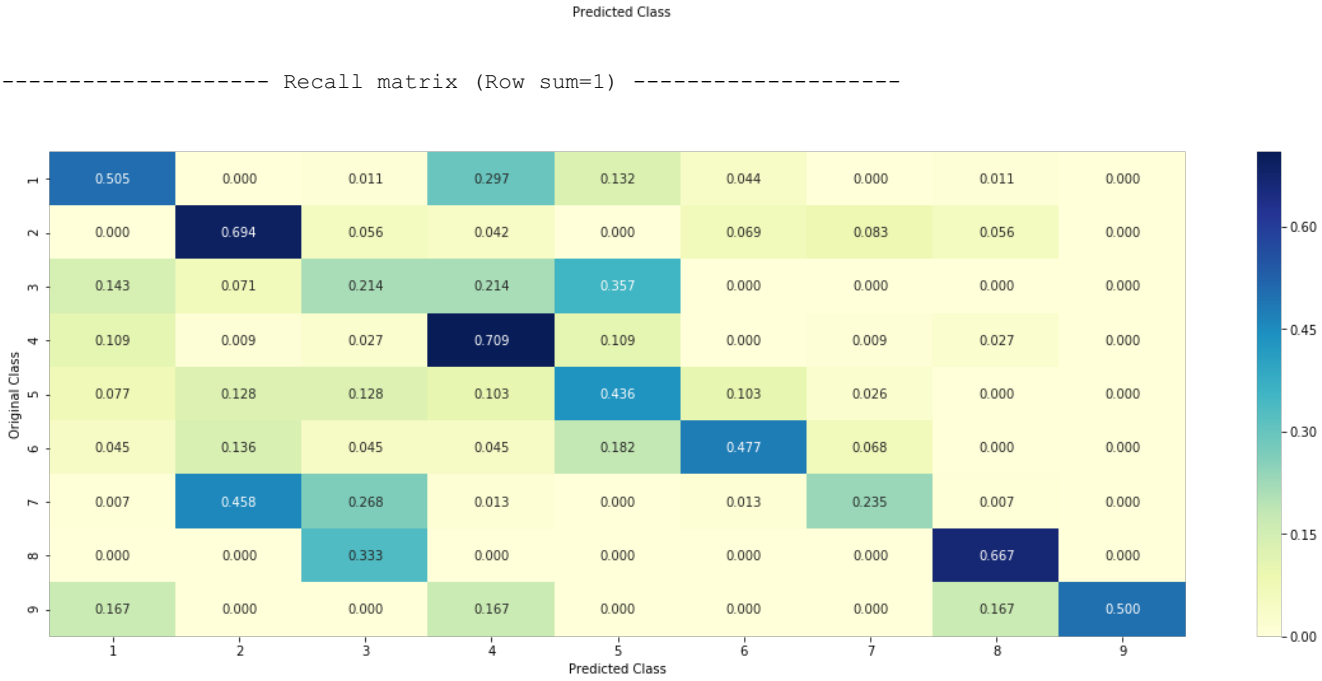
In [0]:

```
# --------------------------------
# default parameters
# sklearn.ensemble.RandomForestClassifier(n_estimators=10, criterion='gini', max_depth=None, min_s
amples_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-fores
t-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='auto',random_state=42)
predict_and_plot_confusion_matrix(train_x_responseCoding, train_y,cv_x_responseCoding,cv_y, clf)
```

```
Log loss : 1.4177501184246677
Number of mis-classified points : 0.518796992481203
-------------------- Confusion matrix --------------------
```


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





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

4.5.5. Feature Importance

4.5.5.1. Correctly Classified point

In [0]:

```
clf = RandomForestClassifier(n_estimators=alpha[int(best_alpha/4)], criterion='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 = 1
no_feature = 27
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 : 2
Predicted Class Probabilities: [[0.0143 0.5044 0.1471 0.0191 0.0245 0.065  0.1724 0.039  0.0142]]
Actual Class : 7
--------------------------------------------------
Variation is important feature
Variation is important feature
Variation is important feature
Variation is important feature
Text is important feature
Variation is important feature
Gene is important feature
Variation is important feature
Text is important feature
Text is important feature
Text is important feature
Gene is important feature
Text is important feature
```

```
Gene is important feature
Variation is important feature
Text is important feature
Gene is important feature
Gene is important feature
Gene is important feature
Variation is important feature
Variation is important feature
Text is important feature
Text is important feature
Gene is important feature
Text is important feature
Gene is important feature
Gene is important feature
```

4.5.5.2. Incorrectly Classified point

In [0]:

```
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("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: [[0.0281 0.2006 0.203  0.0857 0.0626 0.0906 0.2249 0.0676 0.0369]]
Actual Class : 7
--------------------------------------------------
Variation is important feature
Variation is important feature
Variation is important feature
Variation is important feature
Text is important feature
Variation is important feature
Gene is important feature
Variation is important feature
Text is important feature
Text is important feature
Text is important feature
Gene is important feature
Text is important feature
Gene is important feature
Variation is important feature
Text is important feature
Gene is important feature
Gene is important feature
Gene is important feature
Variation is important feature
Variation is important feature
Text is important feature
Text is important feature
Gene is important feature
Text is important feature
Gene is important feature
Gene is important feature
```

4.7 Stack the models

4.7.1 testing with hyper parameter tuning

In [0]:

```
# 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_i
ter=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-in
tuition-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, t
ol=0.001,
# cache_size=200, class_weight=None, verbose=False, max_iter=-1, decision_function_shape='ovr', ra
ndom_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_s
amples_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-fores
t-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=1, 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=0.001)
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_onehot
Coding))))
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_p
robas=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, sc
lf.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
```

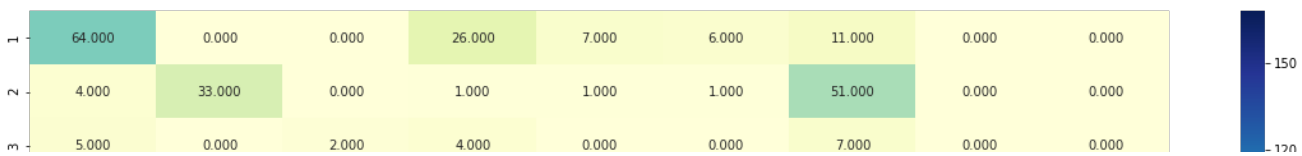
```
Logistic Regression :  Log Loss: 1.24
Support vector machines : Log Loss: 1.72
Naive Bayes : Log Loss: 1.37
--------------------------------------------------
Stacking Classifer : for the value of alpha: 0.000100 Log Loss: 2.179
Stacking Classifer : for the value of alpha: 0.001000 Log Loss: 2.049
Stacking Classifer : for the value of alpha: 0.010000 Log Loss: 1.577
Stacking Classifer : for the value of alpha: 0.100000 Log Loss: 1.224
Stacking Classifer : for the value of alpha: 1.000000 Log Loss: 1.366
Stacking Classifer : for the value of alpha: 10.000000 Log Loss: 1.690
```

4.7.2 testing the model with the best hyper parameters

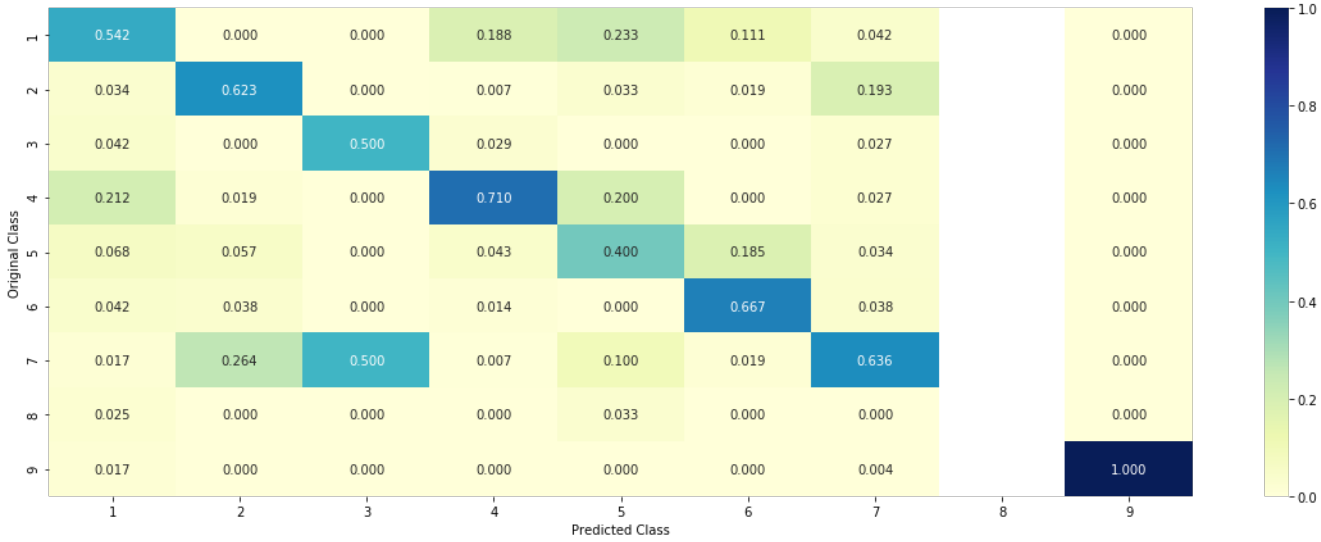
In [0]:

```
lr = LogisticRegression(C=0.1)
sclf = StackingClassifier(classifiers=[sig_clf1, sig_clf2, sig_clf3], meta_classifier=lr, use_proba
s=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.predict(test_x_onehotCoding)-
test_y))/test_y.shape[0])
plot_confusion_matrix(test_y=test_y, predict_y=sclf.predict(test_x_onehotCoding))
```

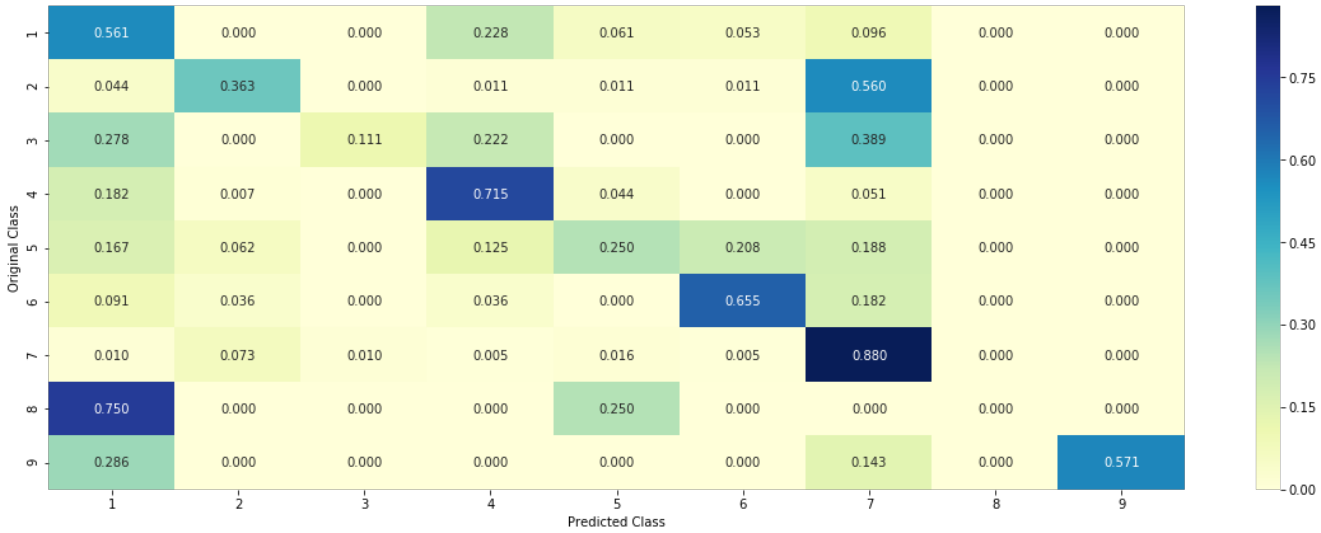
```
Log loss (train) on the stacking classifier : 0.6760284396805781
Log loss (CV) on the stacking classifier : 1.2243084610674686
Log loss (test) on the stacking classifier : 1.1562525475350196
Number of missclassified point : 0.37293233082706767
-------------------- Confusion matrix --------------------
```




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




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



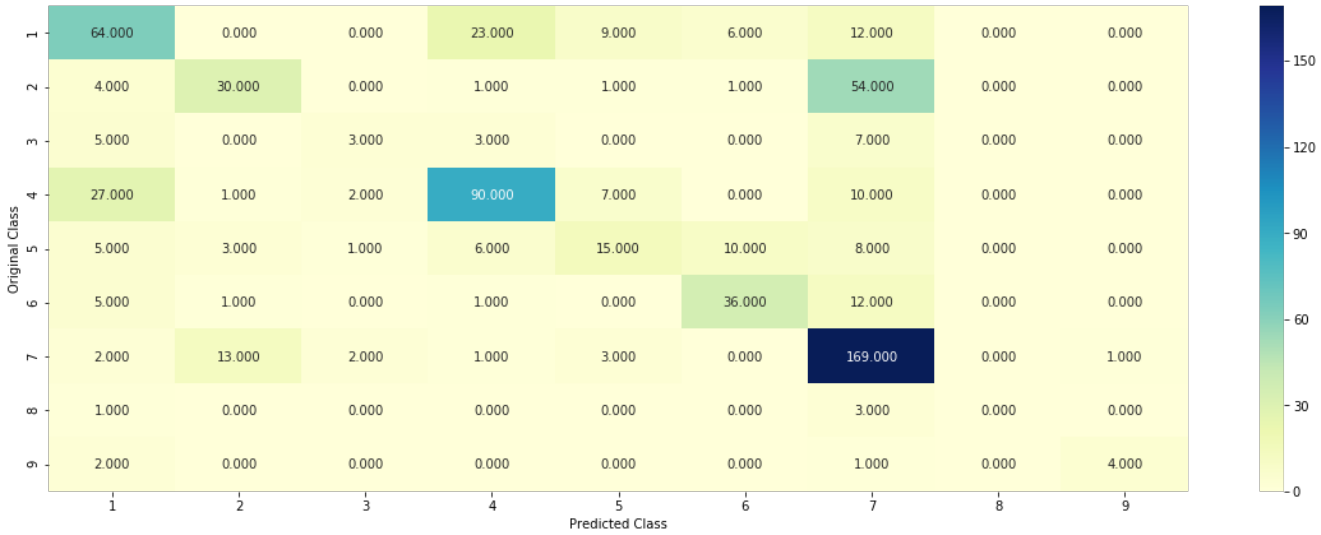
4.7.3 Maximum Voting classifier

In [0]:

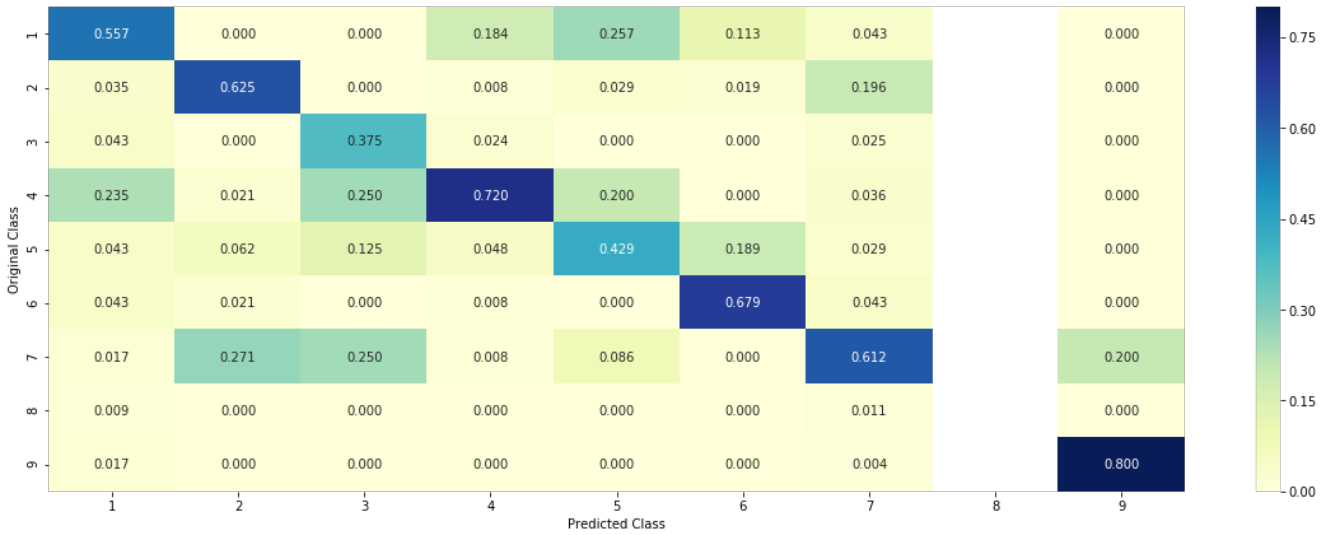
```
#Refer:http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.VotingClassifier.html
from sklearn.ensemble import VotingClassifier
vclf = VotingClassifier(estimators=[('lr', sig_clf1), ('svc', sig_clf2), ('rf', sig_clf3)], voting=
'soft')
vclf.fit(train_x_onehotCoding, train_y)
print("Log loss (train) on the VotingClassifier :", log_loss(train_y,
vclf.predict_proba(train_x_onehotCoding)))
```

```
print("Log loss (CV) on the VotingClassifier :", log_loss(cv_y,
vclf.predict_proba(cv_x_onehotCoding)))
print("Log loss (test) on the VotingClassifier :", log_loss(test_y,
vclf.predict_proba(test_x_onehotCoding)))
print("Number of missclassified point :", np.count_nonzero((vclf.predict(test_x_onehotCoding)-
test_y))/test_y.shape[0])
plot_confusion_matrix(test_y=test_y, predict_y=vclf.predict(test_x_onehotCoding))
```

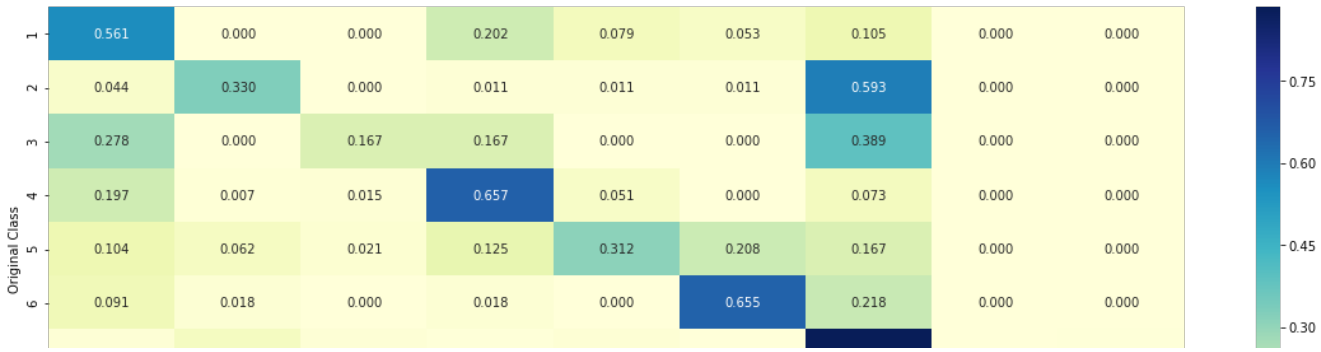
```
Log loss (train) on the VotingClassifier : 0.9407598679043604
Log loss (CV) on the VotingClassifier : 1.2835402100341697
Log loss (test) on the VotingClassifier : 1.223278167176945
Number of missclassified point : 0.3819548872180451
-------------------- Confusion matrix --------------------
```

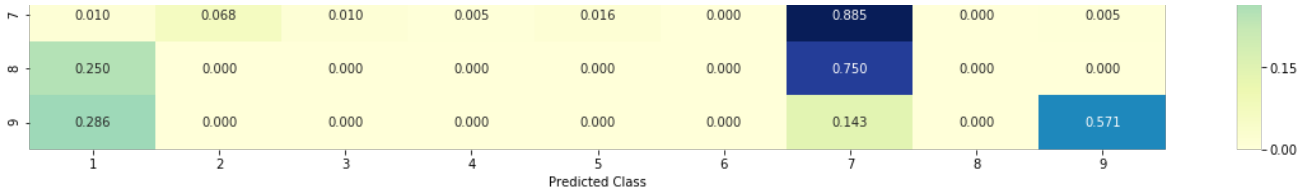



```
-------------------- Precision matrix (Columm Sum=1) --------------------
```




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





5. Assignments

1. Apply All the models with tf-idf features (Replace CountVectorizer with tfidfVectorizer and run the same cells)
2. Instead of using all the words in the dataset, use only the top 1000 words based of tf-idf values
3. Apply Logistic regression with CountVectorizer Features, including both unigrams and bigrams
4. Try any of the feature engineering techniques discussed in the course to reduce the CV and test log-loss to a value less than 1.0

5.1 Using Tfidf as vectorizer for text data and building models on top of that

In [0]:

```
# building a TFIDF with all the words that occured minimum 3 times in train data
vectorizer = TfidfVectorizer(min_df=3)
train_text_feature_tfidf = vectorizer.fit_transform(train_df['TEXT'])
# getting all the feature names (words)
train_text_features= vectorizer.get_feature_names()

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

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


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

```
Total number of unique words in train data : 52690
```

In [0]:

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

# dict_list[i] is build on i'th  class text data
# total_dict is buid on whole training text data
total_dict = extract_dictionary_paddle(train_df)


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]:

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

# we use the same vectorizer that was trained on train data
```

```
test_text_feature_tfidf = vectorizer.transform(test_df['TEXT'])
# don't forget to normalize every feature
test_text_feature_tfidf = normalize(test_text_feature_tfidf, axis=0)

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

Stacking the vectors together

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)
)


train_x_tfidf = hstack((train_gene_var_onehotCoding, train_text_feature_tfidf)).tocsr()

test_x_tfidf = hstack((test_gene_var_onehotCoding, test_text_feature_tfidf)).tocsr()

cv_x_tfidf = hstack((cv_gene_var_onehotCoding, cv_text_feature_tfidf)).tocsr()
```

In [0]:

```
train_y = np.array(list(train_df['Class']))
test_y = np.array(list(test_df['Class']))
cv_y = np.array(list(cv_df['Class']))
```

In [0]:

```
print("TFIDF features :")
print("(number of data points * number of features) in train data = ", train_x_tfidf.shape)
print("(number of data points * number of features) in test data = ", test_x_tfidf.shape)
print("(number of data points * number of features) in cross validation data =", cv_x_tfidf.shape)
```

```
TFIDF features :
(number of data points * number of features) in train data =  (2124, 54877)
(number of data points * number of features) in test data =  (665, 54877)
(number of data points * number of features) in cross validation data = (532, 54877)
```

5.1.1 Base Line Model

5.1.1.1 Naive Bayes

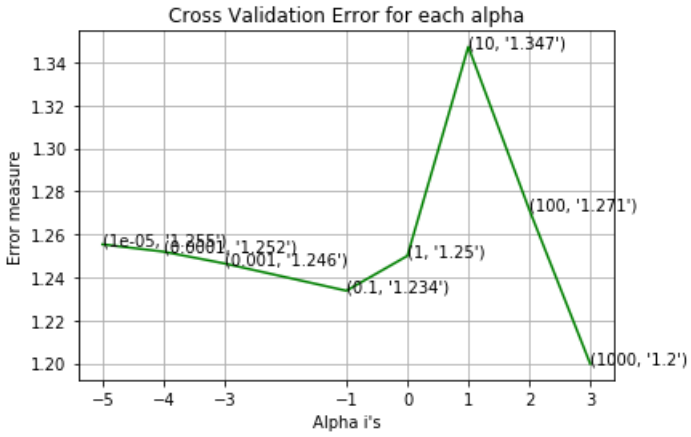
Hyper parameter tuning

In [0]:

```
# 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, method='sigmoid', cv=3)
#
# some of the methods of CalibratedClassifierCV()
# fit(X, y[, sample_weight]) Fit the calibrated model
# get_params([deep]) Get parameters for this estimator.
# predict(X) Predict the target of new samples.
# predict_proba(X) Posterior probabilities of classification
# ----------------------------
# video link: 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_tfidf, train_y)
    sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
    sig_clf.fit(train_x_tfidf, train_y)
    sig_clf_probs = sig_clf.predict_proba(cv_x_tfidf)
    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_array[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_tfidf, train_y)
sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
sig_clf.fit(train_x_tfidf, train_y)


predict_y = sig_clf.predict_proba(train_x_tfidf)
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_tfidf)
print('For values of best alpha = ', alpha[best_alpha], "The cross validation log loss is:",log_lo
ss(y_cv, predict_y, labels=clf.classes_, eps=1e-15))
predict_y = sig_clf.predict_proba(test_x_tfidf)
print('For values of best alpha = ', alpha[best_alpha], "The test log loss is:",log_loss(y_test, p
redict_y, labels=clf.classes_, eps=1e-15))
```

```
for alpha = 1e-05
Log Loss : 1.2553047660264145
for alpha = 0.0001
Log Loss : 1.2519146667597718
for alpha = 0.001
Log Loss : 1.246416404620149
for alpha = 0.1
Log Loss : 1.233732355726079
for alpha = 1
Log Loss : 1.2499437852342876
for alpha = 10
Log Loss : 1.3472263276784744
for alpha = 100
Log Loss : 1.2714401162904636
for alpha = 1000
Log Loss : 1.1997256499141002
```



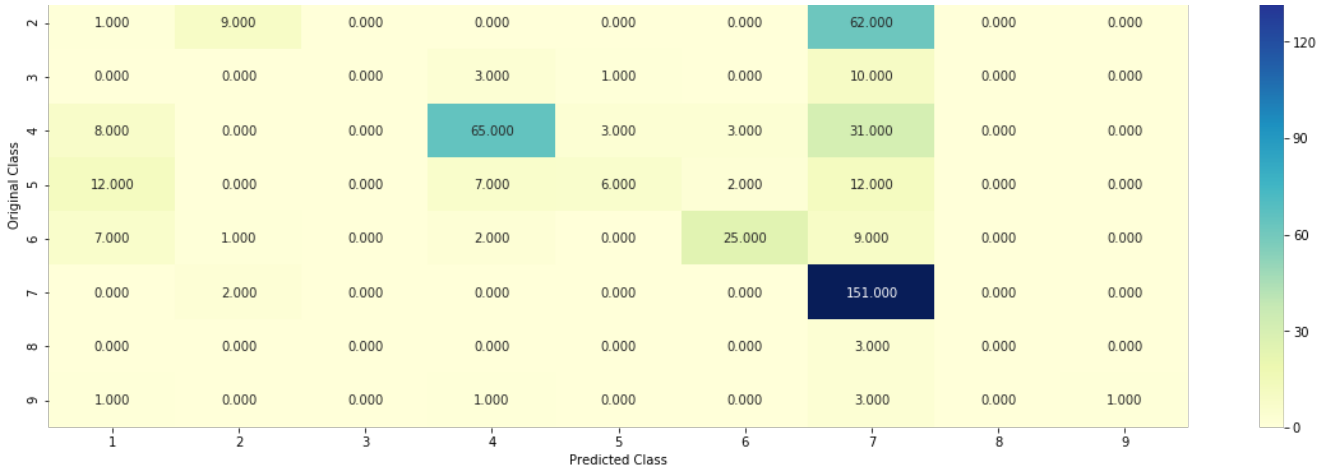

```
For values of best alpha =  1000 The train log loss is: 0.9308385991633955
For values of best alpha =  1000 The cross validation log loss is: 1.1997256499141002
For values of best alpha =  1000 The test log loss is: 1.2582013135537664
```

Testing the model with best hyper paramters

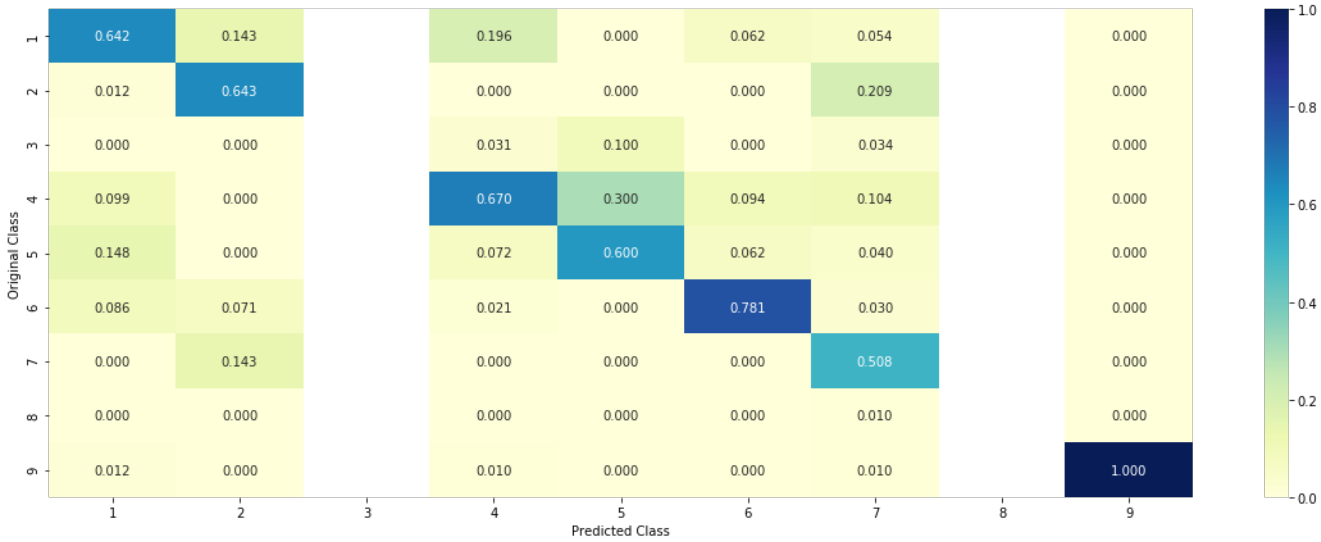
In [0]:

```
# 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, method='sigmoid', cv=3)
#
# some of the methods of CalibratedClassifierCV()
# fit(X, y[, sample_weight]) Fit the calibrated model
# get_params([deep]) Get parameters for this estimator.
# predict(X) Predict the target of new samples.
# predict_proba(X) Posterior probabilities of classification
# ----------------------------

clf = MultinomialNB(alpha=alpha[best_alpha])
clf.fit(train_x_tfidf, train_y)
sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
sig_clf.fit(train_x_tfidf, train_y)
sig_clf_probs = sig_clf.predict_proba(cv_x_tfidf)
# to avoid rounding error while multiplying probabilites we use log-probability estimates
print("Log Loss :",log_loss(cv_y, sig_clf_probs))
print("Number of missclassified point :", np.count_nonzero((sig_clf.predict(cv_x_tfidf)- cv_y))/cv
_y.shape[0])
plot_confusion_matrix(cv_y, sig_clf.predict(cv_x_tfidf.toarray()))
```

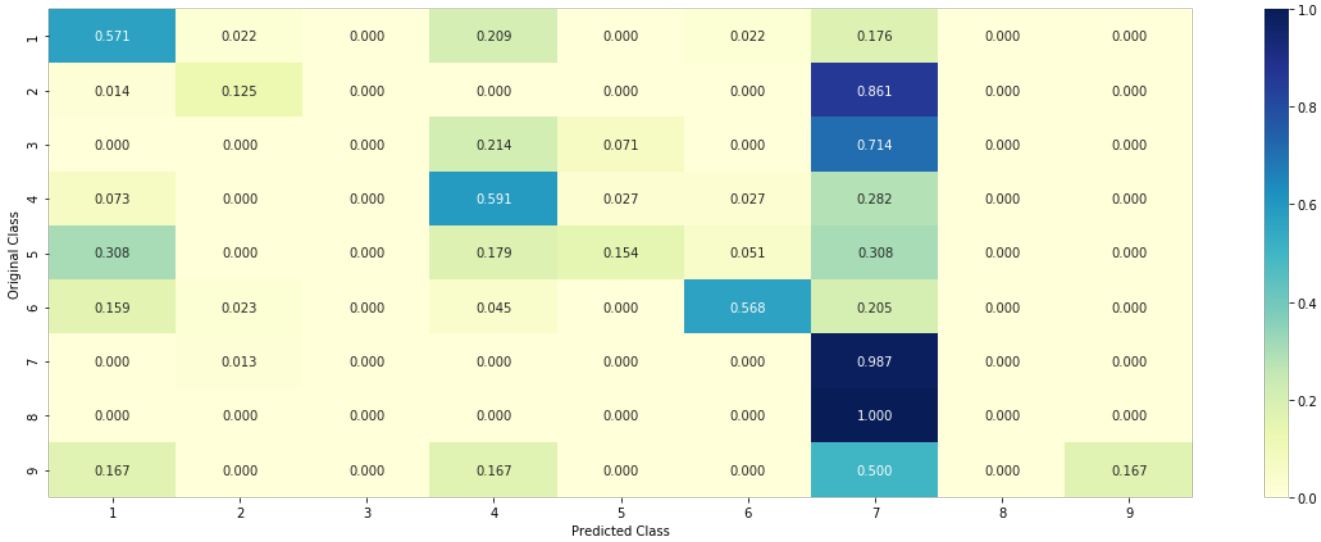
```
Log Loss : 1.1997256499141002
Number of missclassified point : 0.41917293233082711
-------------------- Confusion matrix --------------------
```



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



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



Feature Importance, Correctly classified point

In [0]:

```
test_point_index = 8
no_feature = 100
predicted_cls = sig_clf.predict(test_x_tfidf[test_point_index])
print("Predicted Class :", predicted_cls[0])
```

```
print("Predicted Class Probabilities:",
np.round(sig_clf.predict_proba(test_x_tfidf[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],
test_df['TEXT'].iloc[test_point_index],test_df['Gene'].iloc[test_point_index],test_df['Variation']
.iloc[test_point_index], no_feature)
```

```
Predicted Class : 7
Predicted Class Probabilities: [[3.880e-02 3.600e-02 4.200e-03 5.140e-02 2.100e-02 1.480e-02 8.311
e-01
  2.100e-03 7.000e-04]]
Actual Class : 7
--------------------------------------------------
15 Text feature [cells] present in test data point [True]
16 Text feature [activation] present in test data point [True]
17 Text feature [downstream] present in test data point [True]
18 Text feature [activated] present in test data point [True]
19 Text feature [kinase] present in test data point [True]
20 Text feature [factor] present in test data point [True]
21 Text feature [cell] present in test data point [True]
22 Text feature [contrast] present in test data point [True]
23 Text feature [shown] present in test data point [True]
24 Text feature [phosphorylation] present in test data point [True]
25 Text feature [presence] present in test data point [True]
26 Text feature [also] present in test data point [True]
27 Text feature [inhibitor] present in test data point [True]
28 Text feature [expressing] present in test data point [True]
29 Text feature [recently] present in test data point [True]
30 Text feature [growth] present in test data point [True]
31 Text feature [signaling] present in test data point [True]
32 Text feature [however] present in test data point [True]
34 Text feature [increased] present in test data point [True]
35 Text feature [suggest] present in test data point [True]
36 Text feature [treated] present in test data point [True]
37 Text feature [found] present in test data point [True]
39 Text feature [compared] present in test data point [True]
40 Text feature [addition] present in test data point [True]
41 Text feature [10] present in test data point [True]
42 Text feature [similar] present in test data point [True]
43 Text feature [mutations] present in test data point [True]
44 Text feature [independent] present in test data point [True]
45 Text feature [showed] present in test data point [True]
46 Text feature [interestingly] present in test data point [True]
47 Text feature [previously] present in test data point [True]
48 Text feature [treatment] present in test data point [True]
49 Text feature [enhanced] present in test data point [True]
50 Text feature [well] present in test data point [True]
52 Text feature [potential] present in test data point [True]
53 Text feature [3b] present in test data point [True]
54 Text feature [sensitive] present in test data point [True]
55 Text feature [tyrosine] present in test data point [True]
56 Text feature [pathways] present in test data point [True]
57 Text feature [mechanism] present in test data point [True]
58 Text feature [activating] present in test data point [True]
59 Text feature [constitutive] present in test data point [True]
60 Text feature [may] present in test data point [True]
61 Text feature [observed] present in test data point [True]
62 Text feature [consistent] present in test data point [True]
63 Text feature [inhibitors] present in test data point [True]
64 Text feature [various] present in test data point [True]
65 Text feature [activate] present in test data point [True]
66 Text feature [figure] present in test data point [True]
70 Text feature [furthermore] present in test data point [True]
71 Text feature [fig] present in test data point [True]
73 Text feature [including] present in test data point [True]
74 Text feature [mutant] present in test data point [True]
75 Text feature [demonstrated] present in test data point [True]
76 Text feature [reported] present in test data point [True]
77 Text feature [constitutively] present in test data point [True]
79 Text feature [mutation] present in test data point [True]
80 Text feature [described] present in test data point [True]
82 Text feature [approximately] present in test data point [True]
83 Text feature [inhibition] present in test data point [True]
84 Text feature [increase] present in test data point [True]
87 Text feature [induced] present in test data point [True]
```

```
88 Text feature [using] present in test data point [True]
89 Text feature [without] present in test data point [True]
91 Text feature [two] present in test data point [True]
93 Text feature [report] present in test data point [True]
94 Text feature [due] present in test data point [True]
95 Text feature [role] present in test data point [True]
96 Text feature [recent] present in test data point [True]
97 Text feature [identified] present in test data point [True]
98 Text feature [suggesting] present in test data point [True]
99 Text feature [oncogenic] present in test data point [True]
Out of the top  100  features  72 are present in query point
```

Feature Importance, Incorrectly classified point

In [0]:

```
test_point_index = 1
no_feature = 100
predicted_cls = sig_clf.predict(test_x_tfidf[test_point_index])
print("Predicted Class :", predicted_cls[0])
print("Predicted Class Probabilities:",
np.round(sig_clf.predict_proba(test_x_tfidf[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],
test_df['TEXT'].iloc[test_point_index],test_df['Gene'].iloc[test_point_index],test_df['Variation']
.iloc[test_point_index], no_feature)
```

```
Predicted Class : 7
Predicted Class Probabilities: [[0.1774 0.0882 0.01   0.1081 0.0359 0.026  0.5487 0.0035 0.0023]]
Actual Class : 1
--------------------------------------------------
15 Text feature [cells] present in test data point [True]
16 Text feature [activation] present in test data point [True]
17 Text feature [downstream] present in test data point [True]
18 Text feature [activated] present in test data point [True]
19 Text feature [kinase] present in test data point [True]
20 Text feature [factor] present in test data point [True]
21 Text feature [cell] present in test data point [True]
22 Text feature [contrast] present in test data point [True]
23 Text feature [shown] present in test data point [True]
24 Text feature [phosphorylation] present in test data point [True]
25 Text feature [presence] present in test data point [True]
26 Text feature [also] present in test data point [True]
27 Text feature [inhibitor] present in test data point [True]
28 Text feature [expressing] present in test data point [True]
29 Text feature [recently] present in test data point [True]
30 Text feature [growth] present in test data point [True]
31 Text feature [signaling] present in test data point [True]
32 Text feature [however] present in test data point [True]
34 Text feature [increased] present in test data point [True]
35 Text feature [suggest] present in test data point [True]
36 Text feature [treated] present in test data point [True]
37 Text feature [found] present in test data point [True]
38 Text feature [higher] present in test data point [True]
39 Text feature [compared] present in test data point [True]
40 Text feature [addition] present in test data point [True]
41 Text feature [10] present in test data point [True]
42 Text feature [similar] present in test data point [True]
43 Text feature [mutations] present in test data point [True]
44 Text feature [independent] present in test data point [True]
45 Text feature [showed] present in test data point [True]
47 Text feature [previously] present in test data point [True]
48 Text feature [treatment] present in test data point [True]
50 Text feature [well] present in test data point [True]
51 Text feature [1a] present in test data point [True]
52 Text feature [potential] present in test data point [True]
53 Text feature [3b] present in test data point [True]
54 Text feature [sensitive] present in test data point [True]
55 Text feature [tyrosine] present in test data point [True]
56 Text feature [pathways] present in test data point [True]
57 Text feature [mechanism] present in test data point [True]
58 Text feature [activating] present in test data point [True]
```

```
59 Text feature [constitutive] present in test data point [True]
60 Text feature [may] present in test data point [True]
61 Text feature [observed] present in test data point [True]
62 Text feature [consistent] present in test data point [True]
63 Text feature [inhibitors] present in test data point [True]
65 Text feature [activate] present in test data point [True]
66 Text feature [figure] present in test data point [True]
70 Text feature [furthermore] present in test data point [True]
72 Text feature [obtained] present in test data point [True]
73 Text feature [including] present in test data point [True]
74 Text feature [mutant] present in test data point [True]
78 Text feature [4a] present in test data point [True]
79 Text feature [mutation] present in test data point [True]
80 Text feature [described] present in test data point [True]
83 Text feature [inhibition] present in test data point [True]
84 Text feature [increase] present in test data point [True]
86 Text feature [respectively] present in test data point [True]
87 Text feature [induced] present in test data point [True]
88 Text feature [using] present in test data point [True]
89 Text feature [without] present in test data point [True]
90 Text feature [followed] present in test data point [True]
91 Text feature [two] present in test data point [True]
92 Text feature [antibodies] present in test data point [True]
94 Text feature [due] present in test data point [True]
95 Text feature [role] present in test data point [True]
97 Text feature [identified] present in test data point [True]
98 Text feature [suggesting] present in test data point [True]
Out of the top  100  features  68 are present in query point
```

5.1.2. Logistic Regression

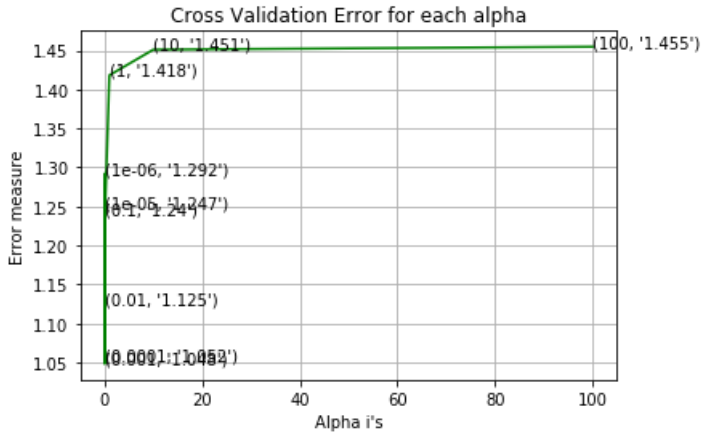
Hyper paramter tuning

In [0]:

```
# 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_i
ter=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-in
tuition-1/
#-------------------------------


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

alpha = [10 ** x for x in range(-6, 3)]
cv_log_error_array = []
for i in alpha:
    print("for alpha =", i)
```

```
    clf = SGDClassifier(class_weight='balanced', alpha=i, penalty='l2', loss='log', random_state=42
)
    clf.fit(train_x_tfidf, train_y)
    sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
    sig_clf.fit(train_x_tfidf, train_y)
    sig_clf_probs = sig_clf.predict_proba(cv_x_tfidf)
    cv_log_error_array.append(log_loss(cv_y, sig_clf_probs, labels=clf.classes_, eps=1e-15))
    # to avoid rounding error while multiplying probabilites we use log-probability estimates
    print("Log Loss :",log_loss(cv_y, sig_clf_probs))

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


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

predict_y = sig_clf.predict_proba(train_x_tfidf)
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_tfidf)
print('For values of best alpha = ', alpha[best_alpha], "The cross validation log loss is:",log_lo
ss(y_cv, predict_y, labels=clf.classes_, eps=1e-15))
predict_y = sig_clf.predict_proba(test_x_tfidf)
print('For values of best alpha = ', alpha[best_alpha], "The test log loss is:",log_loss(y_test, p
redict_y, labels=clf.classes_, eps=1e-15))
```

```
for alpha = 1e-06
Log Loss : 1.2916347394859047
for alpha = 1e-05
Log Loss : 1.2474195572524471
for alpha = 0.0001
Log Loss : 1.0520321911598902
for alpha = 0.001
Log Loss : 1.048172761230156
for alpha = 0.01
Log Loss : 1.1245276639550883
for alpha = 0.1
Log Loss : 1.2399543718924197
for alpha = 1
Log Loss : 1.4182108548992425
for alpha = 10
Log Loss : 1.4510424315901473
for alpha = 100
Log Loss : 1.4547808511639266
```




```
For values of best alpha =  0.001 The train log loss is: 0.5228623386679264
```

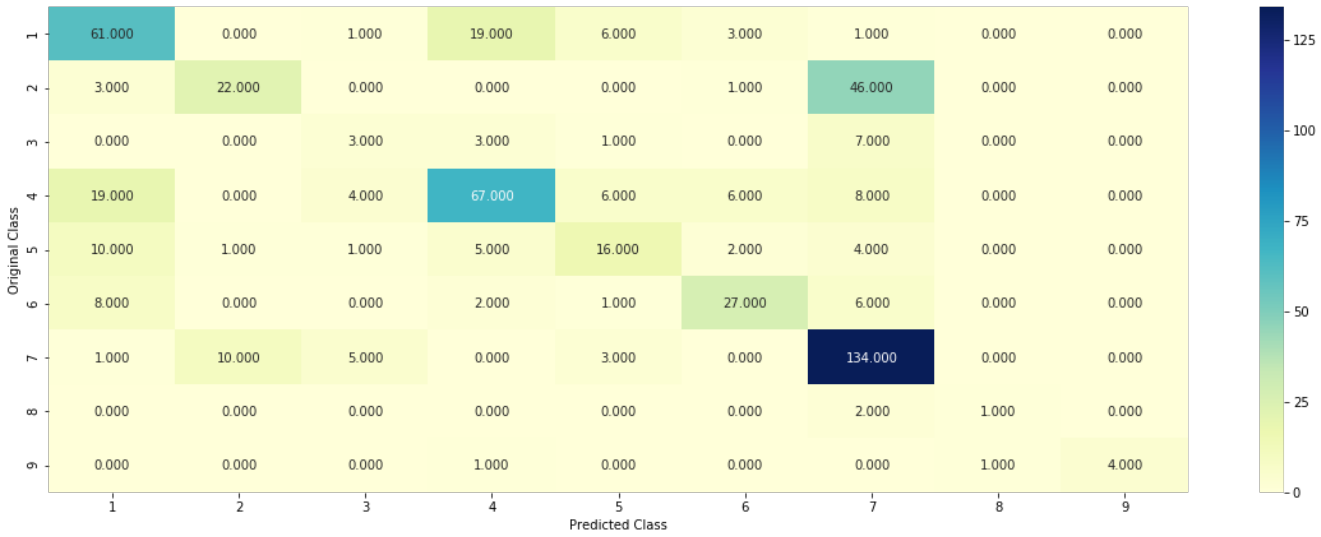
```
For values of best alpha =  0.001 The cross validation log loss is: 1.048172761230156
For values of best alpha =  0.001 The test log loss is: 1.1122438184069563
```

Testing the model with best hyper paramters

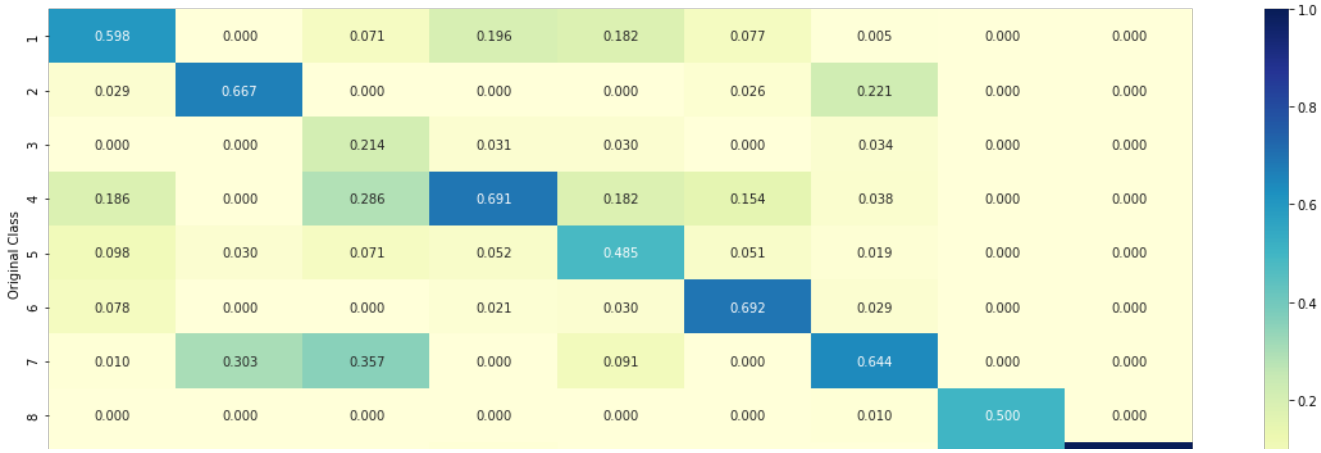
In [0]:

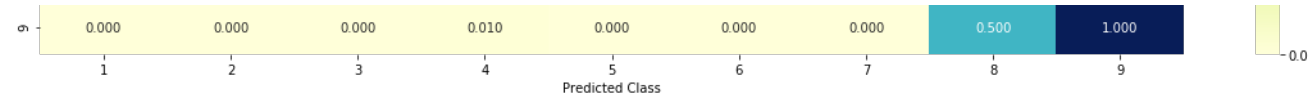
```
# read more about SGDClassifier() at http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDClassifier.html
# ------------------------------
# default parameters
# SGDClassifier(loss='hinge', penalty='l2', alpha=0.0001, l1_ratio=0.15, fit_intercept=True, max_iter=None, tol=None,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

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

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

```
Log loss : 1.048172761230156
Number of mis-classified points : 0.37030075187969924
-------------------- Confusion matrix --------------------
```

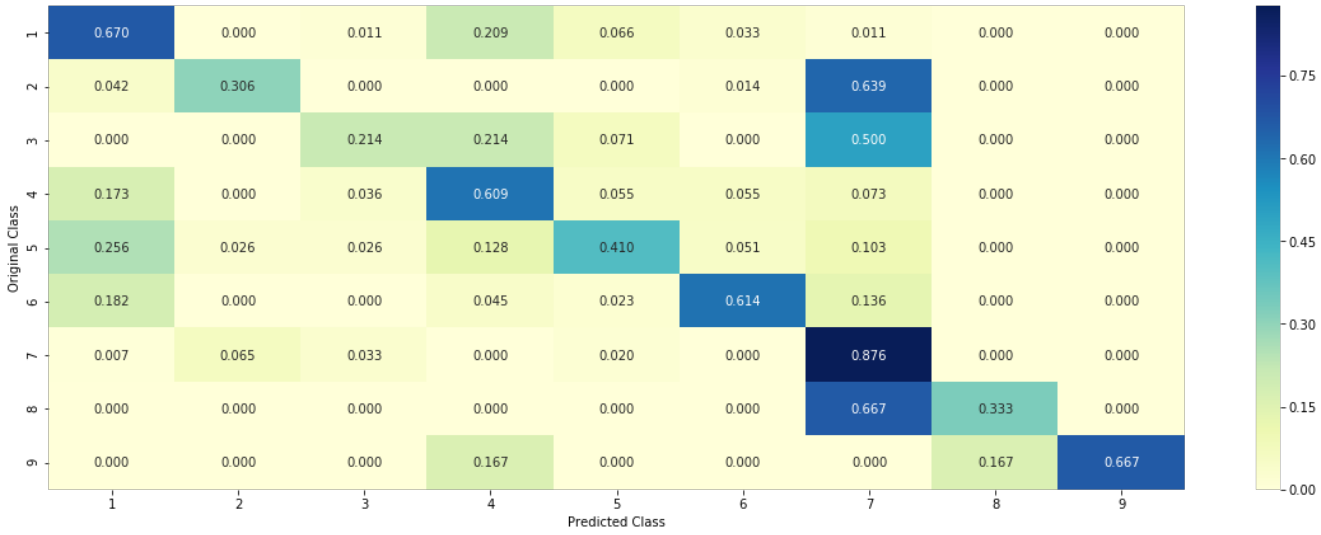



```
-------------------- Precision matrix (Columm Sum=1) --------------------
```






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



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 query point")
    print("-"*50)
    print("The features that are most importent of the ",predicted_cls[0]," class:")
    print (tabulate(tabulte_list, headers=["Index",'Feature name', 'Present or Not']))
```

Correctly Classified point

In [0]:

```
# from tabulate import tabulate
clf = SGDClassifier(class_weight='balanced', alpha=alpha[best_alpha], penalty='l2', loss='log', ran
dom_state=42)
clf.fit(train_x_tfidf,train_y)
test_point_index = 4
no_feature = 500
predicted_cls = sig_clf.predict(test_x_tfidf[test_point_index])
print("Predicted Class :", predicted_cls[0])
print("Predicted Class Probabilities:",
np.round(sig_clf.predict_proba(test_x_tfidf[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],
test_df['TEXT'].iloc[test_point_index],test_df['Gene'].iloc[test_point_index],test_df['Variation']
.iloc[test_point_index], no_feature)
```

```
Predicted Class : 7
Predicted Class Probabilities: [[1.040e-02 2.690e-02 9.000e-04 2.000e-04 1.000e-03 8.000e-04 9.584
e-01
  1.000e-03 5.000e-04]]
Actual Class : 7
--------------------------------------------------
28 Text feature [activated] present in test data point [True]
34 Text feature [activation] present in test data point [True]
48 Text feature [3t3] present in test data point [True]
51 Text feature [downstream] present in test data point [True]
57 Text feature [tyr204] present in test data point [True]
59 Text feature [oncogenes] present in test data point [True]
86 Text feature [tyr] present in test data point [True]
95 Text feature [pharma] present in test data point [True]
98 Text feature [mitogen] present in test data point [True]
102 Text feature [thr202] present in test data point [True]
108 Text feature [activate] present in test data point [True]
115 Text feature [transform] present in test data point [True]
118 Text feature [stat5] present in test data point [True]
129 Text feature [oncoprotein] present in test data point [True]
130 Text feature [mapk] present in test data point [True]
134 Text feature [nude] present in test data point [True]
141 Text feature [activating] present in test data point [True]
153 Text feature [pstat5] present in test data point [True]
155 Text feature [tyrosine] present in test data point [True]
172 Text feature [transformed] present in test data point [True]
187 Text feature [phospho] present in test data point [True]
194 Text feature [glaxosmithkline] present in test data point [True]
195 Text feature [murine] present in test data point [True]
201 Text feature [interleukin] present in test data point [True]
212 Text feature [experienced] present in test data point [True]
220 Text feature [erk1] present in test data point [True]
228 Text feature [phosphorylation] present in test data point [True]
252 Text feature [biopsy] present in test data point [True]
284 Text feature [v600e] present in test data point [True]
303 Text feature [lymphoid] present in test data point [True]
317 Text feature [grew] present in test data point [True]
324 Text feature [enhanced] present in test data point [True]
340 Text feature [ligand] present in test data point [True]
341 Text feature [ba] present in test data point [True]
349 Text feature [soft] present in test data point [True]
360 Text feature [transformation] present in test data point [True]
376 Text feature [f3] present in test data point [True]
378 Text feature [agar] present in test data point [True]
403 Text feature [transforming] present in test data point [True]
408 Text feature [inhibitor] present in test data point [True]
409 Text feature [signaling] present in test data point [True]
427 Text feature [nontransformed] present in test data point [True]
431 Text feature [braf] present in test data point [True]
467 Text feature [pathways] present in test data point [True]
470 Text feature [fabre] present in test data point [True]
472 Text feature [subcutaneously] present in test data point [True]
480 Text feature [extracellular] present in test data point [True]
493 Text feature [transduced] present in test data point [True]
494 Text feature [tyr694] present in test data point [True]
497 Text feature [egf] present in test data point [True]
Out of the top  500  features  50 are present in query point
```

Incorrectly Classified point

In [0]:

```
test_point_index = 1
no_feature = 500
predicted_cls = sig_clf.predict(test_x_tfidf[test_point_index])
print("Predicted Class :", predicted_cls[0])
print("Predicted Class Probabilities:",
np.round(sig_clf.predict_proba(test_x_tfidf[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],
test_df['TEXT'].iloc[test_point_index],test_df['Gene'].iloc[test_point_index],test_df['Variation']
```

```
.iloc[test_point_index], no_feature)
```

```
Predicted Class : 7
Predicted Class Probabilities: [[0.3025 0.1004 0.0112 0.0762 0.0231 0.0202 0.4485 0.0064 0.0115]]
Actual Class : 1
--------------------------------------------------
20 Text feature [constitutive] present in test data point [True]
28 Text feature [activated] present in test data point [True]
33 Text feature [erk] present in test data point [True]
34 Text feature [activation] present in test data point [True]
51 Text feature [downstream] present in test data point [True]
64 Text feature [oncogene] present in test data point [True]
108 Text feature [activate] present in test data point [True]
114 Text feature [technology] present in test data point [True]
141 Text feature [activating] present in test data point [True]
149 Text feature [plx4032] present in test data point [True]
155 Text feature [tyrosine] present in test data point [True]
182 Text feature [receptors] present in test data point [True]
195 Text feature [murine] present in test data point [True]
228 Text feature [phosphorylation] present in test data point [True]
403 Text feature [transforming] present in test data point [True]
405 Text feature [antagonist] present in test data point [True]
408 Text feature [inhibitor] present in test data point [True]
409 Text feature [signaling] present in test data point [True]
431 Text feature [braf] present in test data point [True]
434 Text feature [synergize] present in test data point [True]
465 Text feature [intrinsic] present in test data point [True]
467 Text feature [pathways] present in test data point [True]
497 Text feature [egf] present in test data point [True]
Out of the top  500  features  23 are present in query point
```

5.1.3 Linear Support Vector Machines

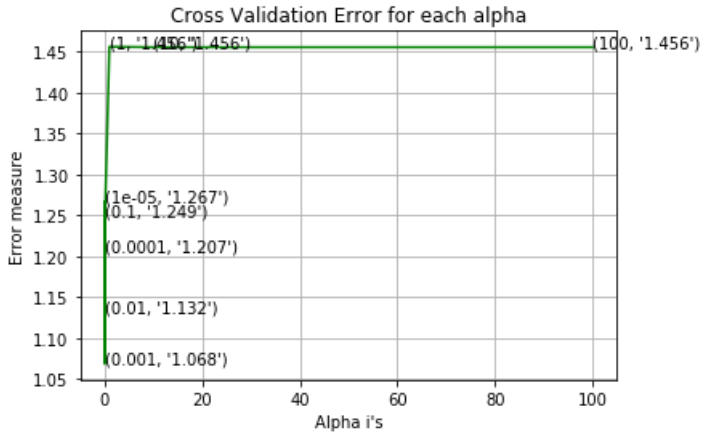
Hyper paramter tuning

In [0]:

```
# 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, t
ol=0.001,
# cache_size=200, class_weight=None, verbose=False, max_iter=-1, decision_function_shape='ovr', ra
ndom_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, method='sigmoid', cv=3)
#
# some of the methods of CalibratedClassifierCV()
# fit(X, y[, sample_weight]) Fit the calibrated model
# get_params([deep]) Get parameters for this estimator.
# predict(X) Predict the target of new samples.
# predict_proba(X) Posterior probabilities of classification
#-------------------------------------
# video link:
#-------------------------------------

alpha = [10 ** x for x in range(-5, 3)]
cv_log_error_array = []
```

```
for i in alpha:
    print("for C =", i)
    clf = SGDClassifier( class_weight='balanced', alpha=i, penalty='l2', loss='hinge', random_state
=42)
    clf.fit(train_x_tfidf, train_y)
    sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
    sig_clf.fit(train_x_tfidf, train_y)
    sig_clf_probs = sig_clf.predict_proba(cv_x_tfidf)
    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 = SVC(C=i,kernel='linear',probability=True, class_weight='balanced')
clf = SGDClassifier(class_weight='balanced', alpha=alpha[best_alpha], penalty='l2', loss='hinge', r
andom_state=42)
clf.fit(train_x_tfidf, train_y)
sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
sig_clf.fit(train_x_tfidf, train_y)

predict_y = sig_clf.predict_proba(train_x_tfidf)
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_tfidf)
print('For values of best alpha = ', alpha[best_alpha], "The cross validation log loss is:",log_lo
ss(y_cv, predict_y, labels=clf.classes_, eps=1e-15))
predict_y = sig_clf.predict_proba(test_x_tfidf)
print('For values of best alpha = ', alpha[best_alpha], "The test log loss is:",log_loss(y_test, p
redict_y, labels=clf.classes_, eps=1e-15))
```

```
for C = 1e-05
Log Loss : 1.2667297127746362
for C = 0.0001
Log Loss : 1.207051411966836
for C = 0.001
Log Loss : 1.0682128589527915
for C = 0.01
Log Loss : 1.1323078823222705
for C = 0.1
Log Loss : 1.2491520697795793
for C = 1
Log Loss : 1.456098449031238
for C = 10
Log Loss : 1.4555186534924047
for C = 100
Log Loss : 1.4555192078504682
```



For values of best alpha = 0.001 The train log loss is: 0.566034857082282

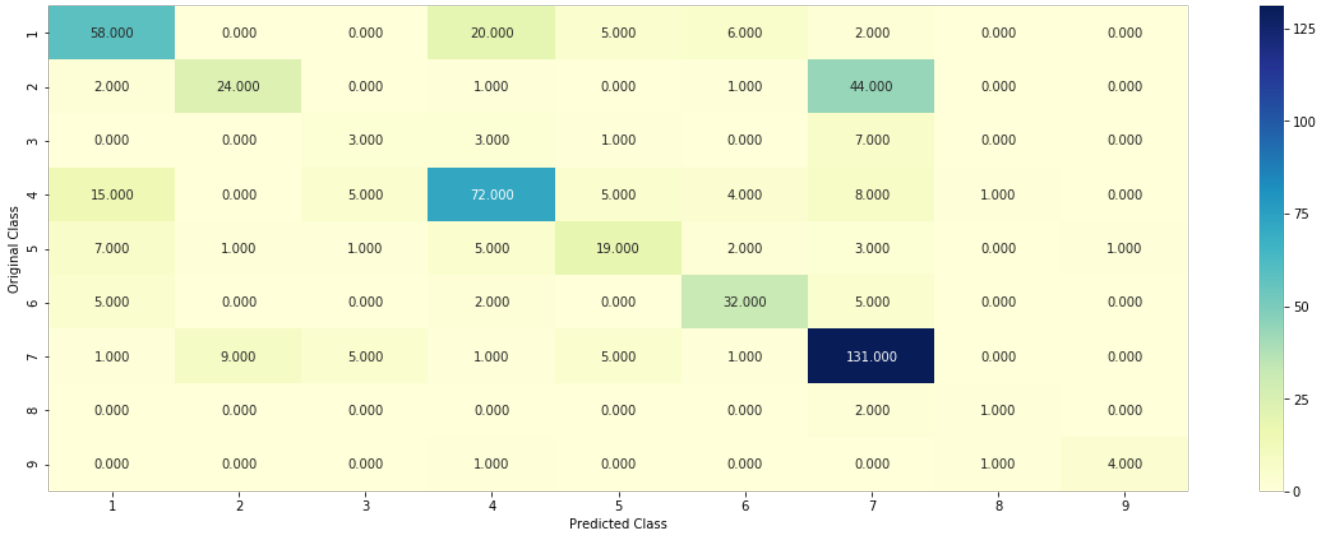
```
For values of best alpha =  0.001 The train log loss is: 0.566034657982282
For values of best alpha =  0.001 The cross validation log loss is: 1.0682128589527915
For values of best alpha =  0.001 The test log loss is: 1.1466278748095227
```

Testing model with best hyper parameters

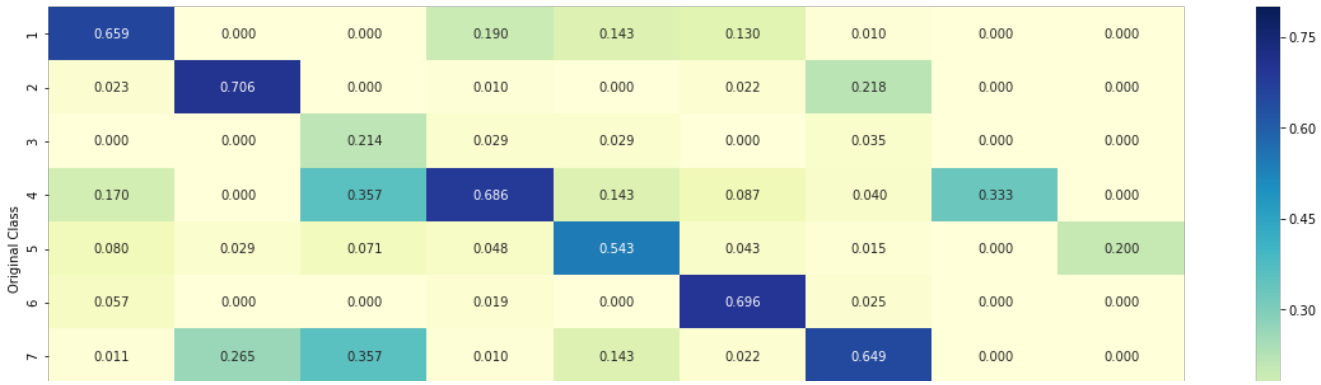
In [0]:

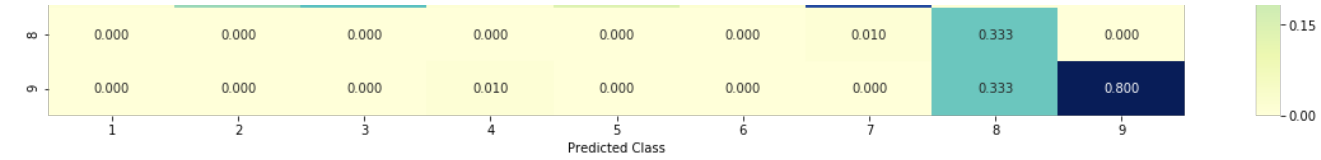
```
# read more about support vector machines with linear kernals here http://scikit-
learn.org/stable/modules/generated/sklearn.svm.SVC.html

# --------------------------------
# default parameters
# SVC(C=1.0, kernel='rbf', degree=3, gamma='auto', coef0=0.0, shrinking=True, probability=False, t
ol=0.001,
# cache_size=200, class_weight=None, verbose=False, max_iter=-1, decision_function_shape='ovr', ra
ndom_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/
# --------------------------------


# 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_tfidf, train_y,cv_x_tfidf,cv_y, clf)
```

```
Log loss : 1.0682128589527915
Number of mis-classified points : 0.3533834586466165
-------------------- Confusion matrix --------------------
```

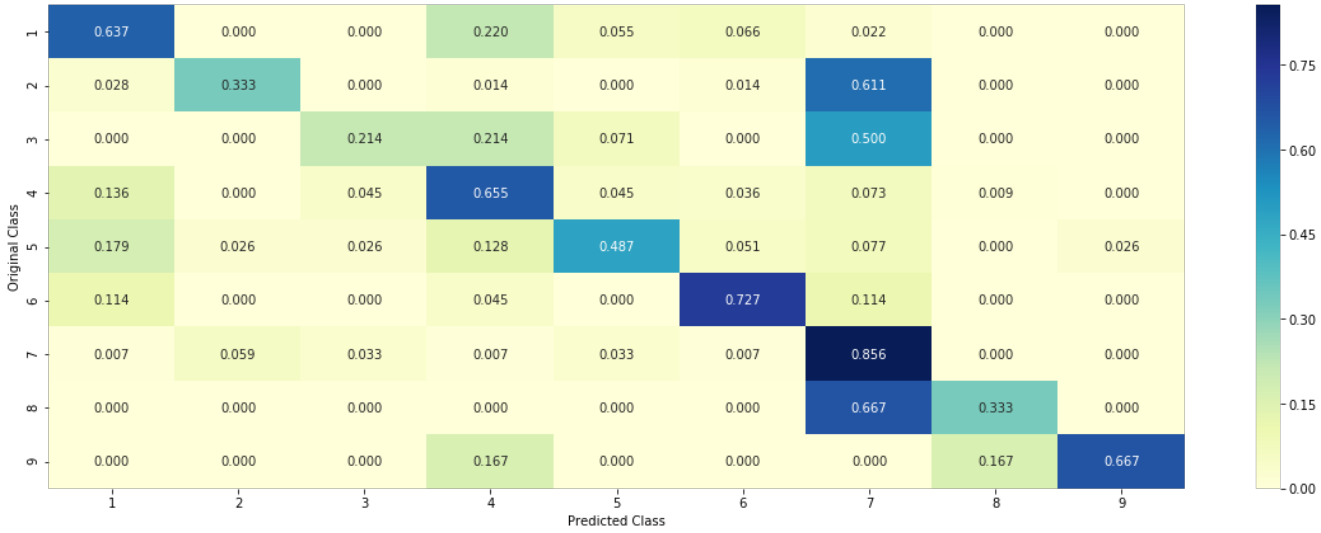



```
-------------------- Precision matrix (Columm Sum=1) --------------------
```





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



Feature Importance

For Correctly classified point

In [0]:

```
clf = SGDClassifier(alpha=alpha[best_alpha], penalty='l2', loss='hinge', random_state=42)
clf.fit(train_x_tfidf,train_y)
test_point_index = 2
# test_point_index = 100
no_feature = 500
predicted_cls = sig_clf.predict(test_x_tfidf[test_point_index])
print("Predicted Class :", predicted_cls[0])
print("Predicted Class Probabilities:",
np.round(sig_clf.predict_proba(test_x_tfidf[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],
test_df['TEXT'].iloc[test_point_index],test_df['Gene'].iloc[test_point_index],test_df['Variation']
.iloc[test_point_index], no_feature)
```

```
Predicted Class : 7
Predicted Class Probabilities: [[0.0438 0.2996 0.0144 0.0654 0.0419 0.0865 0.4314 0.008  0.0089]]
Actual Class : 7
--------------------------------------------------
194 Text feature [activated] present in test data point [True]
201 Text feature [activation] present in test data point [True]
244 Text feature [activate] present in test data point [True]
258 Text feature [oncogene] present in test data point [True]
260 Text feature [erk] present in test data point [True]
263 Text feature [downstream] present in test data point [True]
266 Text feature [transformation] present in test data point [True]
284 Text feature [oncogenes] present in test data point [True]
290 Text feature [phosphorylation] present in test data point [True]
335 Text feature [constitutively] present in test data point [True]
434 Text feature [mitogen] present in test data point [True]
445 Text feature [activates] present in test data point [True]
449 Text feature [expressing] present in test data point [True]
462 Text feature [signaling] present in test data point [True]
497 Text feature [enhanced] present in test data point [True]
Out of the top  500  features  15 are present in query point
```

For Incorrectly classified point

In [0]:

```
test_point_index = 1
no_feature = 500
predicted_cls = sig_clf.predict(test_x_tfidf[test_point_index])
print("Predicted Class :", predicted_cls[0])
print("Predicted Class Probabilities:",
np.round(sig_clf.predict_proba(test_x_tfidf[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],
test_df['TEXT'].iloc[test_point_index],test_df['Gene'].iloc[test_point_index],test_df['Variation']
.iloc[test_point_index], no_feature)
```

```
Predicted Class : 7
Predicted Class Probabilities: [[0.3393 0.0828 0.012  0.0932 0.0224 0.0234 0.41   0.0052 0.0118]]
Actual Class : 1
--------------------------------------------------
161 Text feature [constitutive] present in test data point [True]
194 Text feature [activated] present in test data point [True]
196 Text feature [plx4032] present in test data point [True]
201 Text feature [activation] present in test data point [True]
244 Text feature [activate] present in test data point [True]
258 Text feature [oncogene] present in test data point [True]
260 Text feature [erk] present in test data point [True]
263 Text feature [downstream] present in test data point [True]
290 Text feature [phosphorylation] present in test data point [True]
379 Text feature [technology] present in test data point [True]
435 Text feature [inhibitor] present in test data point [True]
445 Text feature [activates] present in test data point [True]
449 Text feature [expressing] present in test data point [True]
462 Text feature [signaling] present in test data point [True]
Out of the top  500  features  14 are present in query point
```

5.1.4 Random Forest Classifier

Hyper paramter tuning

In [0]:

```
# -------------------------------
# default parameters
# sklearn.ensemble.RandomForestClassifier(n_estimators=10, criterion='gini', max_depth=None, min_s
amples_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-fores
t-and-their-construction-2/
# -------------------------------


# find more about CalibratedClassifierCV here at http://scikit-
learn.org/stable/modules/generated/sklearn.calibration.CalibratedClassifierCV.html
# ----------------------------
# default paramters
```

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

alpha = [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 , class_weight='balanced')
        clf.fit(train_x_tfidf, train_y)
        sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
        sig_clf.fit(train_x_tfidf, train_y)
        sig_clf_probs = sig_clf.predict_proba(cv_x_tfidf)
        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))


best_alpha = np.argmin(cv_log_error_array)
clf = RandomForestClassifier(n_estimators=alpha[int(best_alpha/2)], criterion='gini', max_depth=max
_depth[int(best_alpha%2)], random_state=42, n_jobs=-1)
clf.fit(train_x_tfidf, train_y)
sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
sig_clf.fit(train_x_tfidf, train_y)

predict_y = sig_clf.predict_proba(train_x_tfidf)
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_tfidf)
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_tfidf)
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 : 1.3411231303186424
for n_estimators = 100 and max depth =  10
Log Loss : 1.2058727533292044
for n_estimators = 200 and max depth =  5
Log Loss : 1.2937542646951736
for n_estimators = 200 and max depth =  10
Log Loss : 1.1880720144930734
for n_estimators = 500 and max depth =  5
Log Loss : 1.258453356280692
for n_estimators = 500 and max depth =  10
Log Loss : 1.1686122394928506
for n_estimators = 1000 and max depth =  5
Log Loss : 1.248733110124116
for n_estimators = 1000 and max depth =  10
Log Loss : 1.16529615747781
for n_estimators = 2000 and max depth =  5
Log Loss : 1.23842852623245
for n_estimators = 2000 and max depth =  10
Log Loss : 1.1638683046727905
For values of best estimator =  2000 The train log loss is: 0.6763511826796559
For values of best estimator =  2000 The cross validation log loss is: 1.158495721488525
For values of best estimator =  2000 The test log loss is: 1.1756430020513535
```

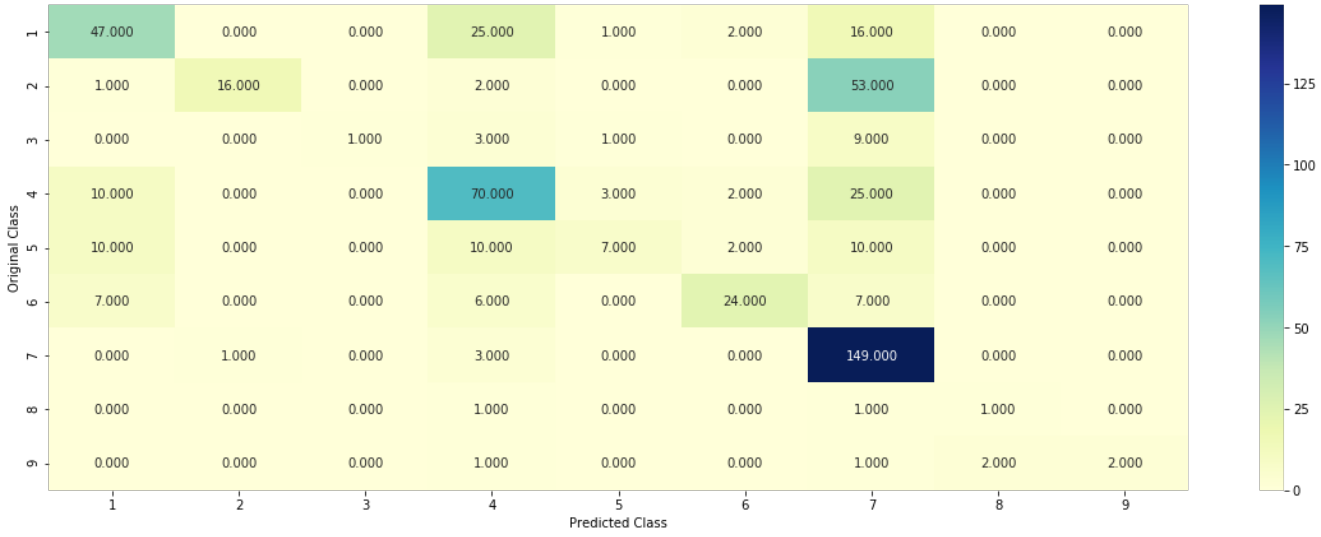
Testing model with best hyper parameters

In [0]:

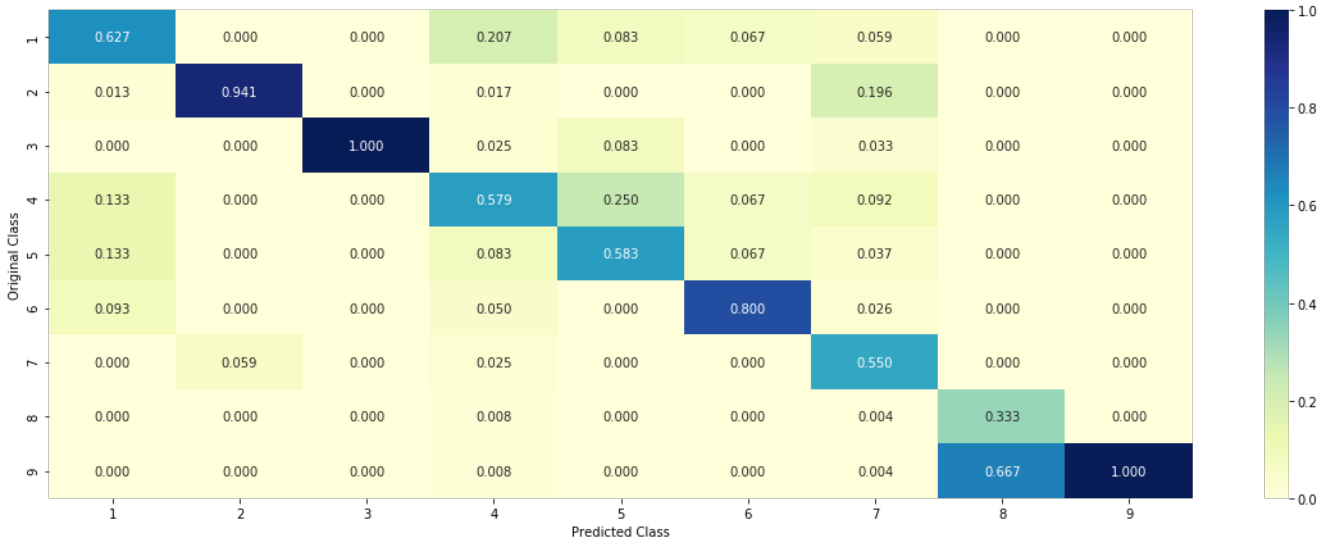
```
# ---------------------------------
# default parameters
```

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

clf = RandomForestClassifier(n_estimators=alpha[int(best_alpha/2)], criterion='gini', max_depth=max_depth[int(best_alpha%2)], random_state=42, n_jobs=-1 , class_weight = 'balanced')
predict_and_plot_confusion_matrix(train_x_tfidf, train_y,cv_x_tfidf,cv_y, clf)
```

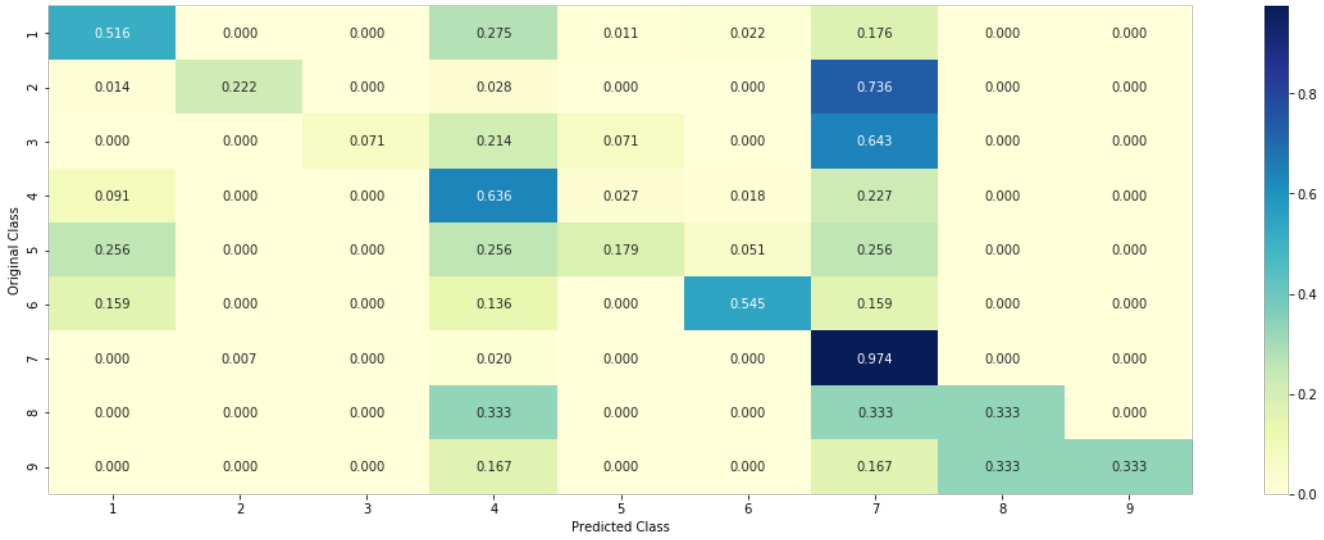
```
Log loss : 1.1638683046727907
Number of mis-classified points : 0.4041353383458647
-------------------- Confusion matrix --------------------
```




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



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



Feature Importance

Incorrectly Classified point

In [0]:

```
# test_point_index = 10
clf = RandomForestClassifier(n_estimators=alpha[int(best_alpha/2)], criterion='gini', max_depth=max
_depth[int(best_alpha%2)], random_state=42, n_jobs=-1 , class_weight = 'balanced')
clf.fit(train_x_tfidf, train_y)
sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
sig_clf.fit(train_x_tfidf, train_y)

test_point_index = 1
no_feature = 100
predicted_cls = sig_clf.predict(test_x_tfidf[test_point_index])
print("Predicted Class :", predicted_cls[0])
print("Predicted Class Probabilities:",
np.round(sig_clf.predict_proba(test_x_tfidf[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], test_df['TEXT'].iloc[test_point_index],test_df['Gene'].
iloc[test_point_index],test_df['Variation'].iloc[test_point_index], no_feature)
```

```
Predicted Class : 7
Predicted Class Probabilities: [[0.0624 0.2034 0.0188 0.0572 0.0446 0.0354 0.5664 0.0059 0.0059]]
Actual Class : 1
--------------------------------------------------
1 Text feature [variants] present in test data point [True]
6 Text feature [kinase] present in test data point [True]
15 Text feature [treatment] present in test data point [True]
20 Text feature [expression] present in test data point [True]
23 Text feature [inhibitors] present in test data point [True]
29 Text feature [functional] present in test data point [True]
30 Text feature [activation] present in test data point [True]
38 Text feature [2009] present in test data point [True]
42 Text feature [activating] present in test data point [True]
44 Text feature [rna] present in test data point [True]
55 Text feature [mesenchymal] present in test data point [True]
58 Text feature [therapy] present in test data point [True]
63 Text feature [fully] present in test data point [True]
70 Text feature [activated] present in test data point [True]
79 Text feature [inhibitor] present in test data point [True]
83 Text feature [predictive] present in test data point [True]
91 Text feature [sequencing] present in test data point [True]
96 Text feature [tyrosine] present in test data point [True]
99 Text feature [classify] present in test data point [True]
Out of the top  100  features  19 are present in query point
```

Correctly Classified point

In [0]:

```
test_point_index = 100
no_feature = 100
predicted_cls = sig_clf.predict(test_x_tfidf[test_point_index])
print("Predicted Class :", predicted_cls[0])
print("Predicted Class Probabilities:",
np.round(sig_clf.predict_proba(test_x_tfidf[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], test_df['TEXT'].iloc[test_point_index],test_df['Gene'].
iloc[test_point_index],test_df['Variation'].iloc[test_point_index], no_feature)
```

```
Predicted Class : 4
Predicted Class Probabilities: [[0.0262 0.0102 0.0607 0.8174 0.0328 0.0225 0.0226 0.0035 0.0041]]
Actual Class : 4
--------------------------------------------------
6 Text feature [kinase] present in test data point [True]
20 Text feature [expression] present in test data point [True]
27 Text feature [predicted] present in test data point [True]
28 Text feature [missense] present in test data point [True]
29 Text feature [functional] present in test data point [True]
30 Text feature [activation] present in test data point [True]
39 Text feature [deleterious] present in test data point [True]
42 Text feature [activating] present in test data point [True]
53 Text feature [alignments] present in test data point [True]
58 Text feature [therapy] present in test data point [True]
63 Text feature [fully] present in test data point [True]
73 Text feature [expected] present in test data point [True]
75 Text feature [likely] present in test data point [True]
83 Text feature [predictive] present in test data point [True]
91 Text feature [sequencing] present in test data point [True]
96 Text feature [tyrosine] present in test data point [True]
Out of the top  100  features  16 are present in query point
```

5.1.5 Stack the models

Testing with hyper parameter tuning

In [0]:

```
# 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_i
ter=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-in
tuition-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, t
ol=0.001,
# cache_size=200, class_weight=None, verbose=False, max_iter=-1, decision_function_shape='ovr', ra
ndom_state=None)
```

```
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_s
amples_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-fores
t-and-their-construction-2/
# --------------------------------


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

clf2 = SGDClassifier(alpha=1, penalty='l2', loss='hinge', class_weight='balanced', random_state=0)
clf2.fit(train_x_tfidf, train_y)
sig_clf2 = CalibratedClassifierCV(clf2, method="sigmoid")


clf3 = MultinomialNB(alpha=0.001)
clf3.fit(train_x_tfidf, train_y)
sig_clf3 = CalibratedClassifierCV(clf3, method="sigmoid")

sig_clf1.fit(train_x_tfidf, train_y)
print("Logistic Regression :  Log Loss: %0.2f" % (log_loss(cv_y, sig_clf1.predict_proba(cv_x_tfidf)
)))
sig_clf2.fit(train_x_tfidf, train_y)
print("Support vector machines : Log Loss: %0.2f" % (log_loss(cv_y,
sig_clf2.predict_proba(cv_x_tfidf))))
sig_clf3.fit(train_x_tfidf, train_y)
print("Naive Bayes : Log Loss: %0.2f" % (log_loss(cv_y, sig_clf3.predict_proba(cv_x_tfidf))))
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_p
robas=True)
    sclf.fit(train_x_tfidf, train_y)
    print("Stacking Classifer : for the value of alpha: %f Log Loss: %0.3f" % (i, log_loss(cv_y, sc
lf.predict_proba(cv_x_tfidf))))
    log_error =log_loss(cv_y, sclf.predict_proba(cv_x_tfidf))
    if best_alpha > log_error:
        best_alpha = log_error
```

```
Logistic Regression :  Log Loss: 1.05
Support vector machines : Log Loss: 1.46
Naive Bayes : Log Loss: 1.25
```

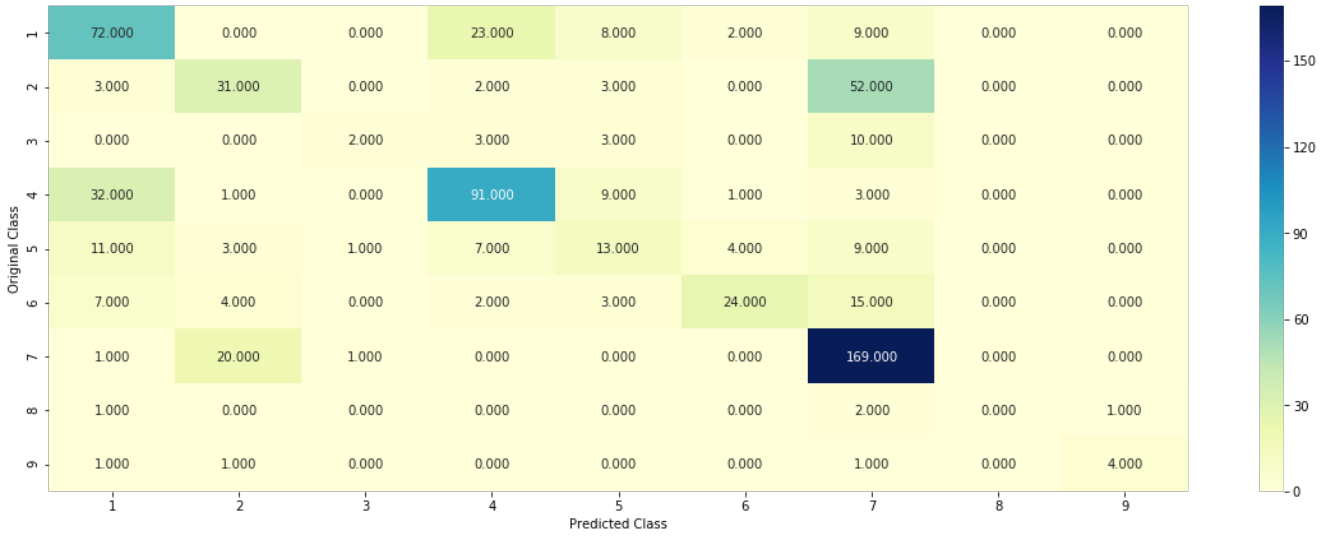
```
--------------------------------------------------
Stacking Classifer : for the value of alpha: 0.000100 Log Loss: 2.178
Stacking Classifer : for the value of alpha: 0.001000 Log Loss: 2.032
Stacking Classifer : for the value of alpha: 0.010000 Log Loss: 1.496
Stacking Classifer : for the value of alpha: 0.100000 Log Loss: 1.113
Stacking Classifer : for the value of alpha: 1.000000 Log Loss: 1.209
Stacking Classifer : for the value of alpha: 10.000000 Log Loss: 1.492
```

Testing the model with the best hyper parameters

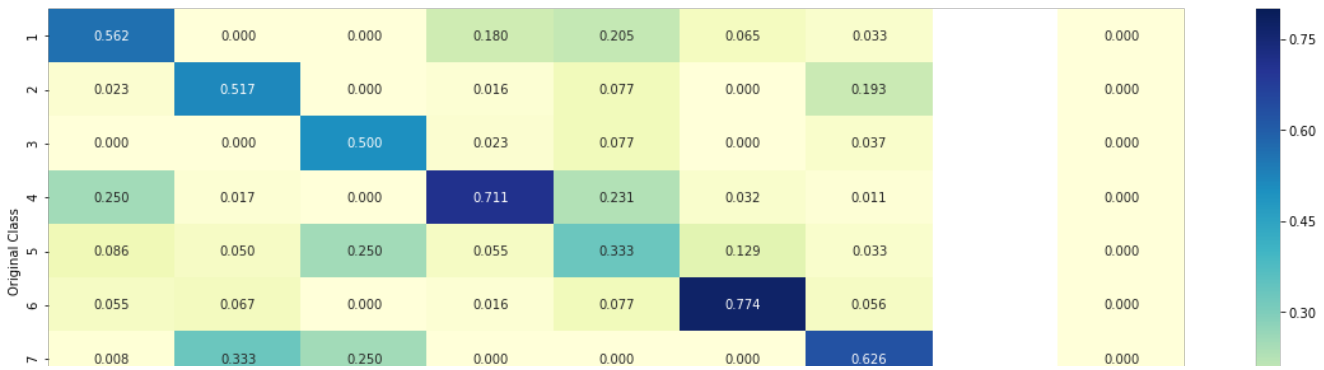
In [0]:

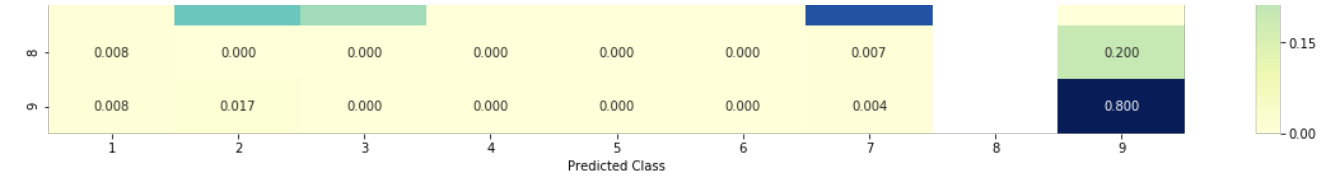
```
lr = LogisticRegression(C=0.1)
sclf = StackingClassifier(classifiers=[sig_clf1, sig_clf2, sig_clf3], meta_classifier=lr, use_proba
s=True)
sclf.fit(train_x_tfidf, train_y)

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

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

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

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

```
Log loss (train) on the stacking classifier : 0.6214005240452861
Log loss (CV) on the stacking classifier : 1.1127593449880218
Log loss (test) on the stacking classifier : 1.1660705613654627
Number of missclassified point : 0.3894736842105263
-------------------- Confusion matrix --------------------
```

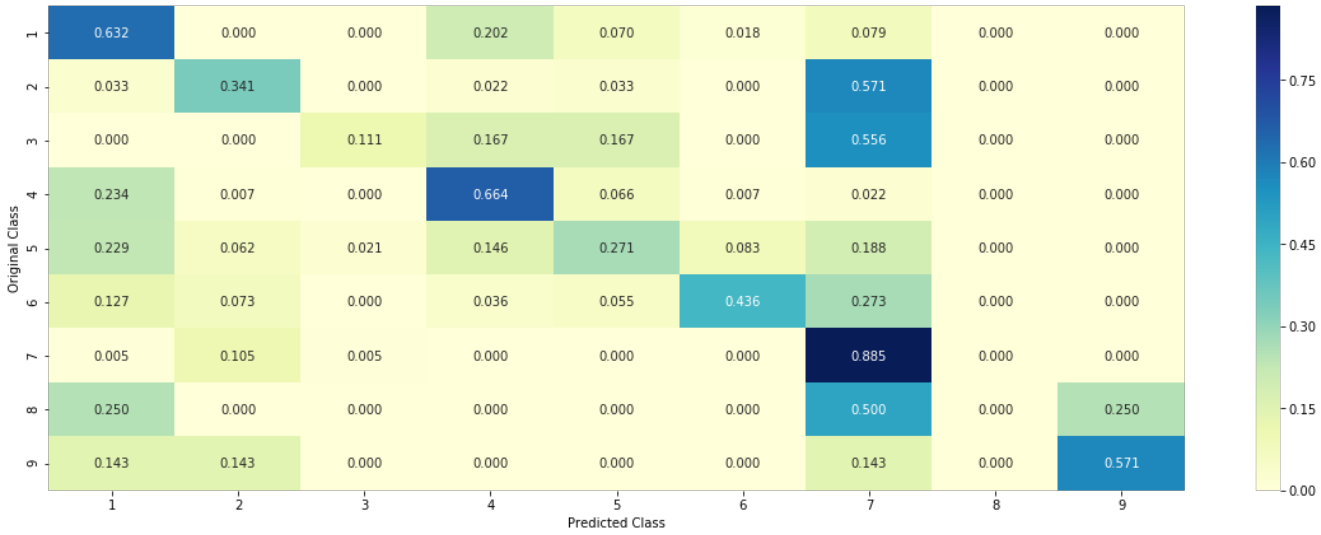



```
-------------------- Precision matrix (Columm Sum=1) --------------------
```






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

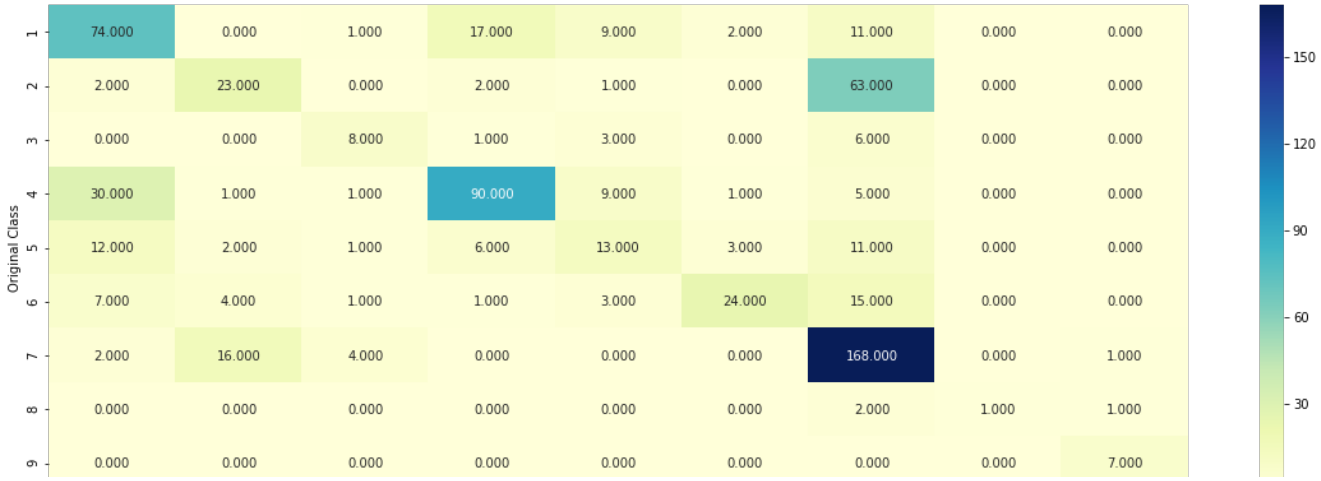


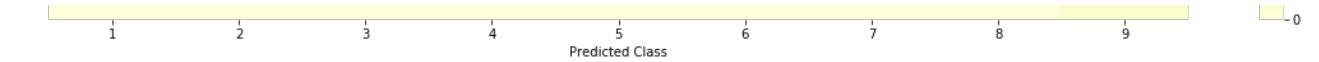
5.1.5.2 Maximum Voting classifier

In [0]:

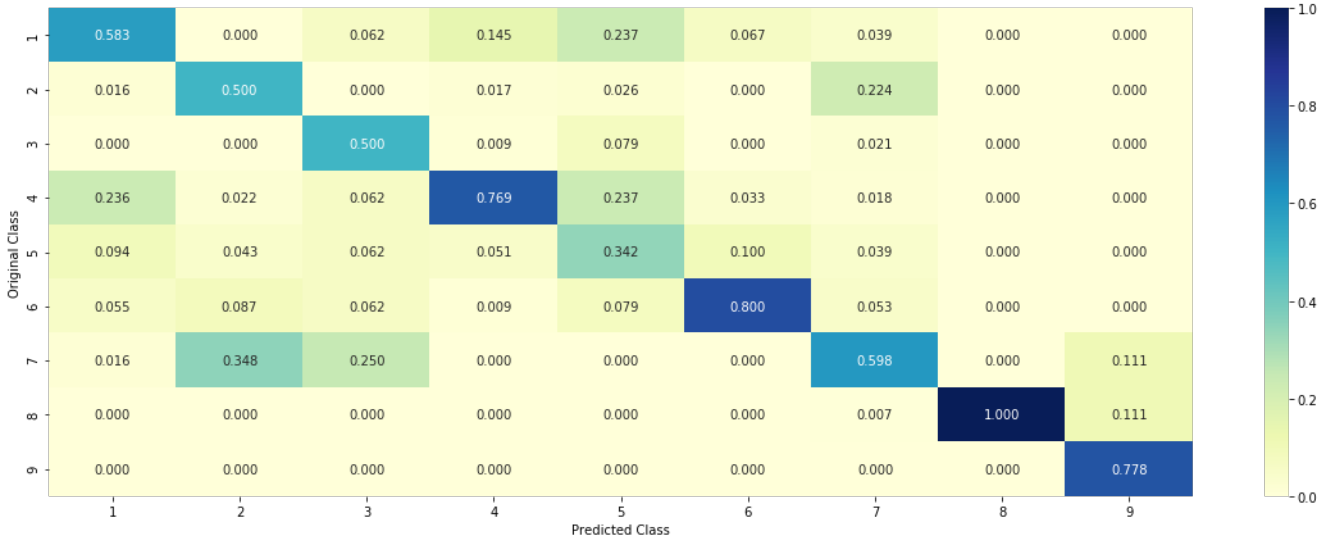
```
#Refer:http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.VotingClassifier.html
from sklearn.ensemble import VotingClassifier
vclf = VotingClassifier(estimators=[('lr', sig_clf1), ('svc', sig_clf2), ('rf', sig_clf3)], voting=
'soft')
vclf.fit(train_x_tfidf, train_y)
print("Log loss (train) on the VotingClassifier :", log_loss(train_y,
vclf.predict_proba(train_x_tfidf)))
print("Log loss (CV) on the VotingClassifier :", log_loss(cv_y, vclf.predict_proba(cv_x_tfidf)))
print("Log loss (test) on the VotingClassifier :", log_loss(test_y,
vclf.predict_proba(test_x_tfidf)))
print("Number of missclassified point :", np.count_nonzero((vclf.predict(test_x_tfidf)- test_y))/t
est_y.shape[0])
plot_confusion_matrix(test_y=test_y, predict_y=vclf.predict(test_x_tfidf))
```

```
Log loss (train) on the VotingClassifier : 0.83703340714461 59
Log loss (CV) on the VotingClassifier : 1.1253005604023416
Log loss (test) on the VotingClassifier : 1.1676448089460476
Number of missclassified point : 0.38646616541353385
-------------------- Confusion matrix --------------------
```

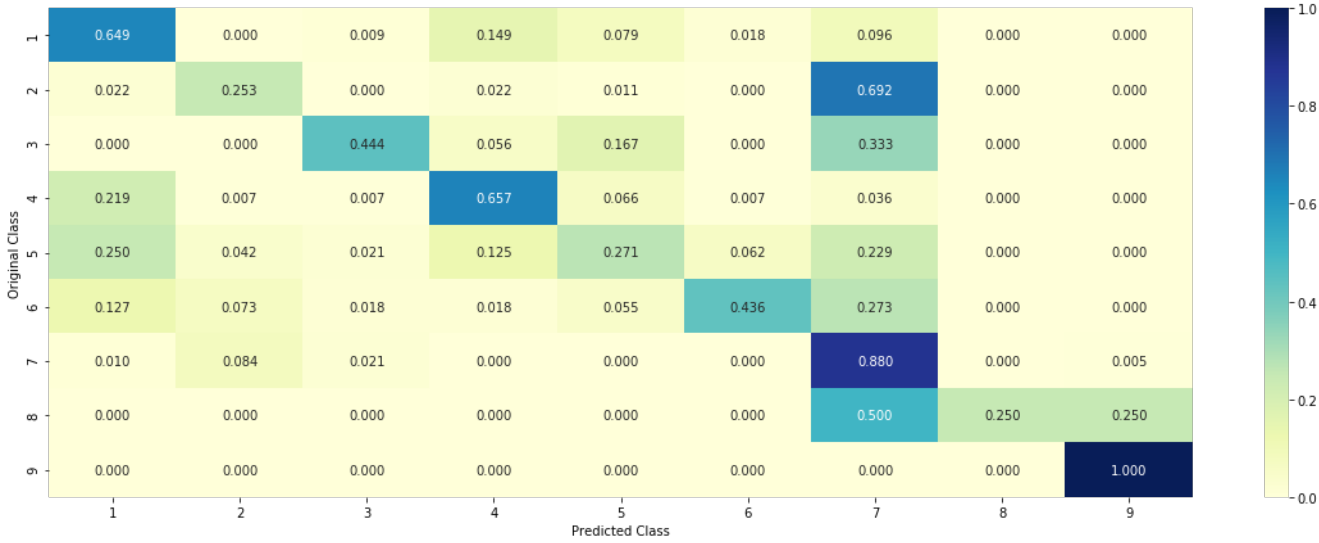





```
-------------------- Precision matrix (Columm Sum=1) --------------------
```




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



5.2 Using top 1000 features of TFIDF vector and building models on top of that

In [0]:

```
from sklearn.feature_selection import SelectKBest
from sklearn.feature_selection import chi2
selection = SelectKBest(chi2, k=1000).fit(train_text_feature_tfidf, y_train)

train_text_feature_tfidf = selection.transform(train_text_feature_tfidf)
test_text_feature_tfidf = selection.transform(test_text_feature_tfidf)
cv_text_feature_tfidf = selection.transform(cv_text_feature_tfidf)

train_text_feature_tfidf = normalize(train_text_feature_tfidf, axis=0)
test_text_feature_tfidf = normalize(test_text_feature_tfidf, axis=0)
cv_text_feature_tfidf = normalize(cv_text_feature_tfidf, axis=0)
```

Stacking the vectors together

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)
)


train_x_tfidf = hstack((train_gene_var_onehotCoding, train_text_feature_tfidf)).tocsr()

test_x_tfidf = hstack((test_gene_var_onehotCoding, test_text_feature_tfidf)).tocsr()

cv_x_tfidf = hstack((cv_gene_var_onehotCoding, cv_text_feature_tfidf)).tocsr()
```

In [0]:

```
train_y = np.array(list(train_df['Class']))
test_y = np.array(list(test_df['Class']))
cv_y = np.array(list(cv_df['Class']))
```

In [0]:

```
print("TFIDF features :")
print("(number of data points * number of features) in train data = ", train_x_tfidf.shape)
print("(number of data points * number of features) in test data = ", test_x_tfidf.shape)
print("(number of data points * number of features) in cross validation data =", cv_x_tfidf.shape)
```

```
TFIDF features :
(number of data points * number of features) in train data =  (2124, 3187)
(number of data points * number of features) in test data =  (665, 3187)
(number of data points * number of features) in cross validation data = (532, 3187)
```

In [0]:

```
#https://towardsdatascience.com/feature-selection-techniques-in-machine-learning-with-python-f24e7
da3f36e
dfscores = pd.DataFrame(selection.scores_)
dfcolumns = pd.DataFrame(train_text_features)
#concat two dataframes for better visualization
featureScores = pd.concat([dfcolumns,dfscores],axis=1)
featureScores.columns = ['Feature names','Score']  #naming the dataframe columns
print(featureScores.nlargest(10,'Score'))  #print 10 best features
```

```
      Feature names       Score
44365         sf3b1  276.741690
49571         u2af2  269.923319
45515   spliceosome  257.609261
52661         zrsr2  253.340321
49568          u2af  250.431015
45129         snrnp  248.398719
27734         k666n  247.477134
33351         n626d  245.333788
52290         y623c  245.333788
5793          abcb7  244.626117
```

In [0]:

```
featureScores.sort_values('Score' , inplace=True , ascending=False)
featureScores.head()
```

Out[0]:

	Feature names	Score
44365	sf3b1	276.741690
49571	u2af2	269.923319
45515	spliceosome	257.609261
52661	zrsr2	253.340321

	Feature names	Score
49568	u2af	250.431015

In [0]:

```
train_text_features = list(featureScores['Feature names'][:1000])
train_text_features[:5]
```

Out[0]:

```
['sf3b1', 'u2af2', 'spliceosome', 'zrsr2', 'u2af']
```

In [0]:

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

# dict_list[i] is build on i'th  class text data
# total_dict is buid on whole training text data
total_dict = extract_dictionary_paddle(train_df)


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)
```

5.2.1 Base Line Model

5.2.1.1 Naive Bayes

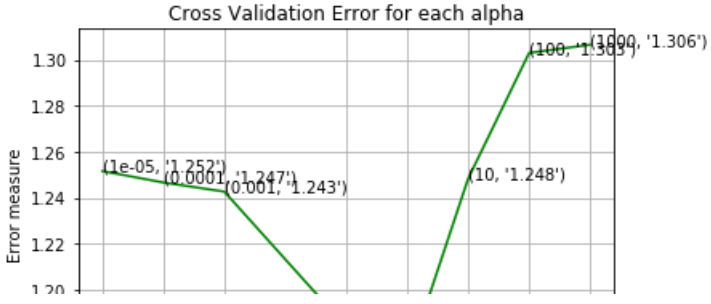
Hyper parameter tuning

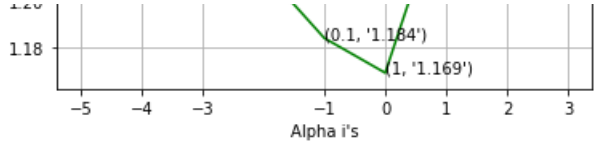
In [0]:

```
# 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, method='sigmoid', cv=3)
#
# some of the methods of CalibratedClassifierCV()
# fit(X, y[, sample_weight]) Fit the calibrated model
# get_params([deep]) Get parameters for this estimator.
# predict(X) Predict the target of new samples.
# predict_proba(X) Posterior probabilities of classification
```

```
# ------------------------------
# video link: 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_tfidf, train_y)
    sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
    sig_clf.fit(train_x_tfidf, train_y)
    sig_clf_probs = sig_clf.predict_proba(cv_x_tfidf)
    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_array[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_tfidf, train_y)
sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
sig_clf.fit(train_x_tfidf, train_y)


predict_y = sig_clf.predict_proba(train_x_tfidf)
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_tfidf)
print('For values of best alpha = ', alpha[best_alpha], "The cross validation log loss is:",log_lo
ss(y_cv, predict_y, labels=clf.classes_, eps=1e-15))
predict_y = sig_clf.predict_proba(test_x_tfidf)
print('For values of best alpha = ', alpha[best_alpha], "The test log loss is:",log_loss(y_test, p
redict_y, labels=clf.classes_, eps=1e-15))
```

```
for alpha = 1e-05
Log Loss : 1.2515771282282728
for alpha = 0.0001
Log Loss : 1.246621804530252
for alpha = 0.001
Log Loss : 1.242663350413691
for alpha = 0.1
Log Loss : 1.1838136544599094
for alpha = 1
Log Loss : 1.1686785389910932
for alpha = 10
Log Loss : 1.2482907123177438
for alpha = 100
Log Loss : 1.3029362186939246
for alpha = 1000
Log Loss : 1.306494692309031
```





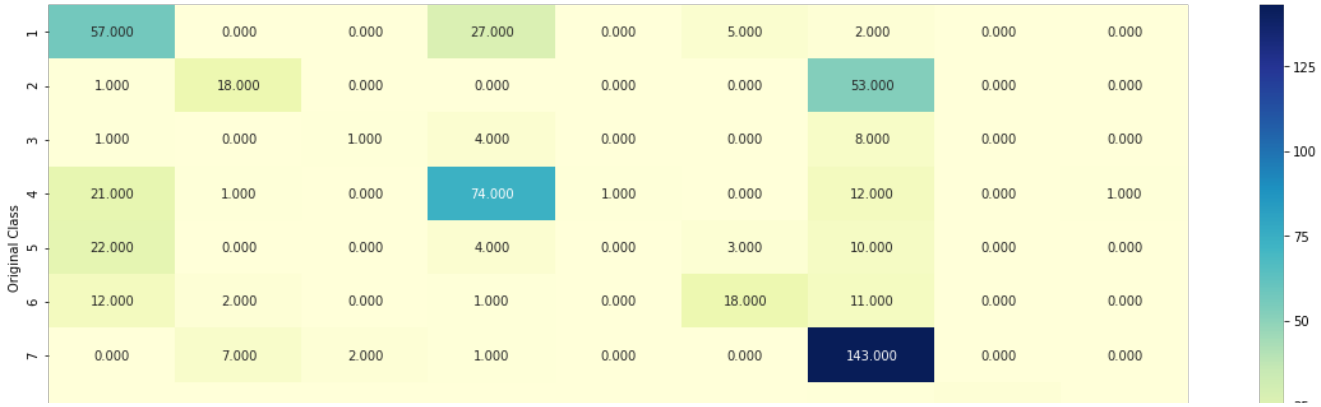

```
For values of best alpha =  1 The train log loss is: 0.8045730815212971
For values of best alpha =  1 The cross validation log loss is: 1.1686785389910932
For values of best alpha =  1 The test log loss is: 1.1732916904620825
```

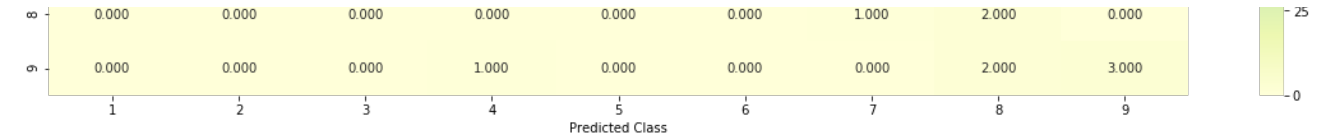
Testing the model with best hyper paramters

In [0]:

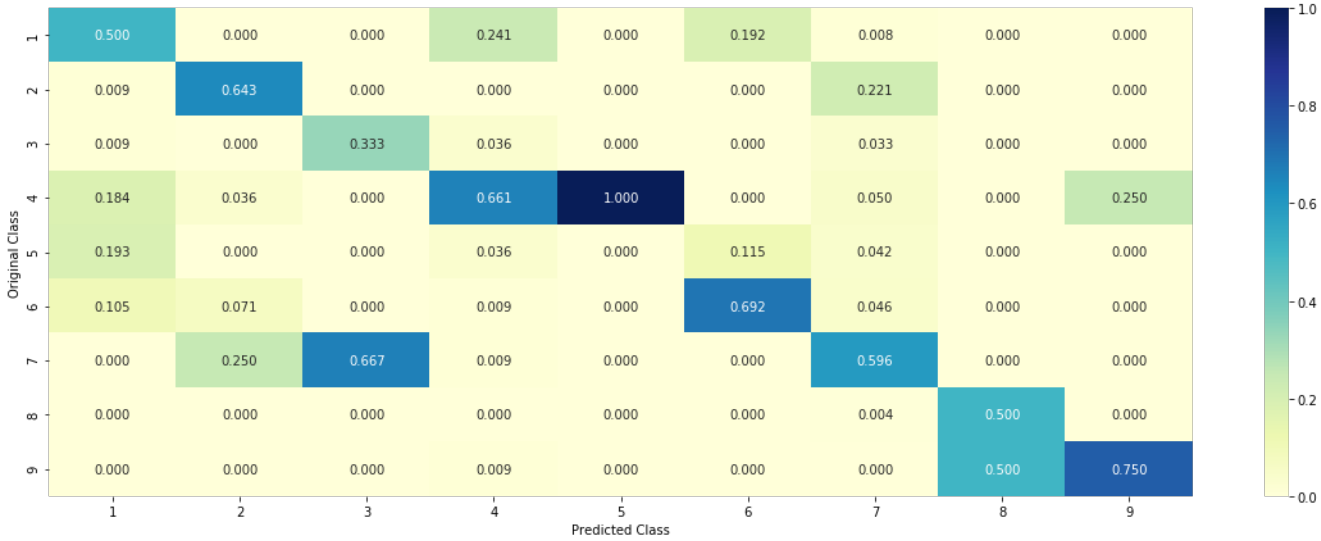
```
# find more about Multinomial Naive base function here http://scikit-
learn.org/stable/modules/generated/sklearn.naive_bayes.MultinomialNB.html
# -------------------------
# default paramters
# sklearn.naive_bayes.MultinomialNB(alpha=1.0, fit_prior=True, class_prior=None)

# 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, method='sigmoid', cv=3)
#
# some of the methods of CalibratedClassifierCV()
# fit(X, y[, sample_weight]) Fit the calibrated model
# get_params([deep]) Get parameters for this estimator.
# predict(X) Predict the target of new samples.
# predict_proba(X) Posterior probabilities of classification
# ----------------------------

clf = MultinomialNB(alpha=alpha[best_alpha])
clf.fit(train_x_tfidf, train_y)
sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
sig_clf.fit(train_x_tfidf, train_y)
sig_clf_probs = sig_clf.predict_proba(cv_x_tfidf)
# to avoid rounding error while multiplying probabilites we use log-probability estimates
print("Log Loss :",log_loss(cv_y, sig_clf_probs))
print("Number of missclassified point :", np.count_nonzero((sig_clf.predict(cv_x_tfidf)- cv_y))/cv
_y.shape[0])
plot_confusion_matrix(cv_y, sig_clf.predict(cv_x_tfidf.toarray()))
```

```
Log Loss : 1.1686785389910932
Number of missclassified point : 0.40601503759398494
-------------------- Confusion matrix --------------------
```

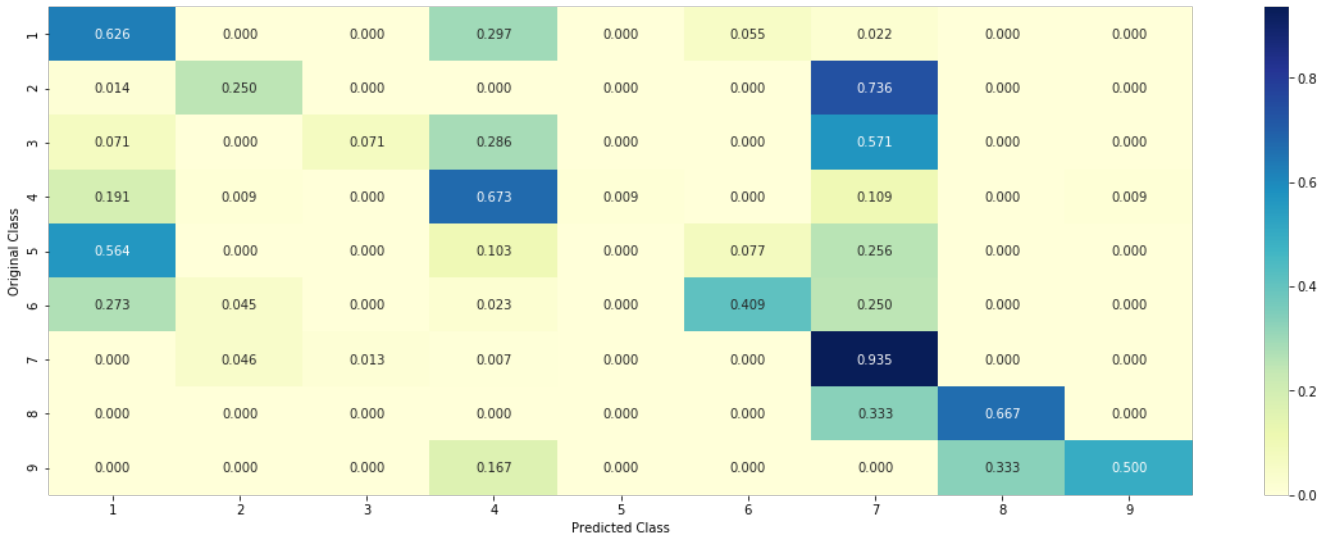




---------- Precision matrix (Columm Sum=1) ----------



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



Feature Importance, Correctly classified point

In [0]:

```
test_point_index = 8
no_feature = 3000
predicted_cls = sig_clf.predict(test_x_tfidf[test_point_index])
print("Predicted Class :", predicted_cls[0])
print("Predicted Class Probabilities:",
np.round(sig_clf.predict_proba(test_x_tfidf[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],
test_df['TEXT'].iloc[test_point_index],test_df['Gene'].iloc[test_point_index],test_df['Variation']
.iloc[test_point_index], no_feature)
```

```
Predicted Class : 7
Predicted Class Probabilities: [[0.0735 0.0943 0.0316 0.0847 0.0596 0.0629 0.5808 0.0063 0.0063]]
Actual Class : 7
```

```
---------------------------------------------------
1010 Text feature [12] present in test data point [True]
1433 Text feature [1172] present in test data point [True]
1483 Text feature [11] present in test data point [True]
1583 Text feature [13] present in test data point [True]
1883 Text feature [100] present in test data point [True]
1884 Text feature [10] present in test data point [True]
1926 Text feature [101] present in test data point [True]
Out of the top  3000  features  7 are present in query point
```

Feature Importance, Incorrectly classified point

In [0]:

```
test_point_index = 5
no_feature = 3000
predicted_cls = sig_clf.predict(test_x_tfidf[test_point_index])
print("Predicted Class :", predicted_cls[0])
print("Predicted Class Probabilities:",
np.round(sig_clf.predict_proba(test_x_tfidf[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],
test_df['TEXT'].iloc[test_point_index],test_df['Gene'].iloc[test_point_index],test_df['Variation']
.iloc[test_point_index], no_feature)
```

```
Predicted Class : 4
Predicted Class Probabilities: [[0.2231 0.0986 0.0335 0.3983 0.0621 0.0653 0.1062 0.0065 0.0064]]
Actual Class : 1
---------------------------------------------------
673 Text feature [12] present in test data point [True]
976 Text feature [13] present in test data point [True]
1223 Text feature [10] present in test data point [True]
1225 Text feature [1000] present in test data point [True]
1260 Text feature [101] present in test data point [True]
1426 Text feature [11] present in test data point [True]
1563 Text feature [01] present in test data point [True]
1723 Text feature [05] present in test data point [True]
Out of the top  3000  features  8 are present in query point
```

5.2.2. Logistic Regression

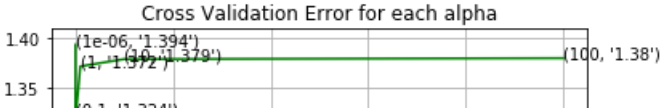
Hyper paramter tuning

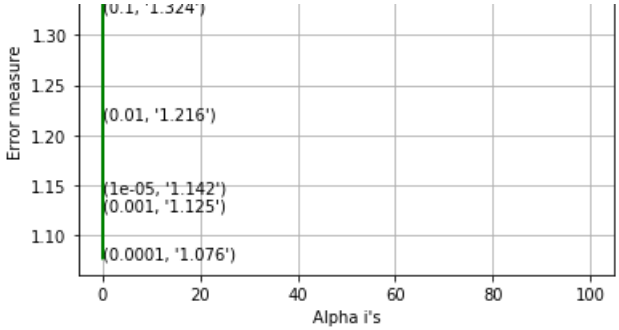
In [0]:

```
# 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_i
ter=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-in
tuition-1/
#-------------------------------


# find more about CalibratedClassifierCV here at http://scikit-
learn.org/stable/modules/generated/sklearn.calibration.CalibratedClassifierCV.html
# ----------------------------
# default paramters
# sklearn.calibration.CalibratedClassifierCV(base_estimator=None, method='sigmoid', cv=3)
```

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

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

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


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

predict_y = sig_clf.predict_proba(train_x_tfidf)
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_tfidf)
print('For values of best alpha = ', alpha[best_alpha], "The cross validation log loss is:",log_lo
ss(y_cv, predict_y, labels=clf.classes_, eps=1e-15))
predict_y = sig_clf.predict_proba(test_x_tfidf)
print('For values of best alpha = ', alpha[best_alpha], "The test log loss is:",log_loss(y_test, p
redict_y, labels=clf.classes_, eps=1e-15))
```

```
for alpha = 1e-06
Log Loss : 1.3935031320143252
for alpha = 1e-05
Log Loss : 1.1421671706992755
for alpha = 0.0001
Log Loss : 1.0759275940084208
for alpha = 0.001
Log Loss : 1.124865901427264
for alpha = 0.01
Log Loss : 1.215652764796414
for alpha = 0.1
Log Loss : 1.3238905707505757
for alpha = 1
Log Loss : 1.371870093518622
for alpha = 10
Log Loss : 1.379106500735401
for alpha = 100
Log Loss : 1.3800598901492096
```





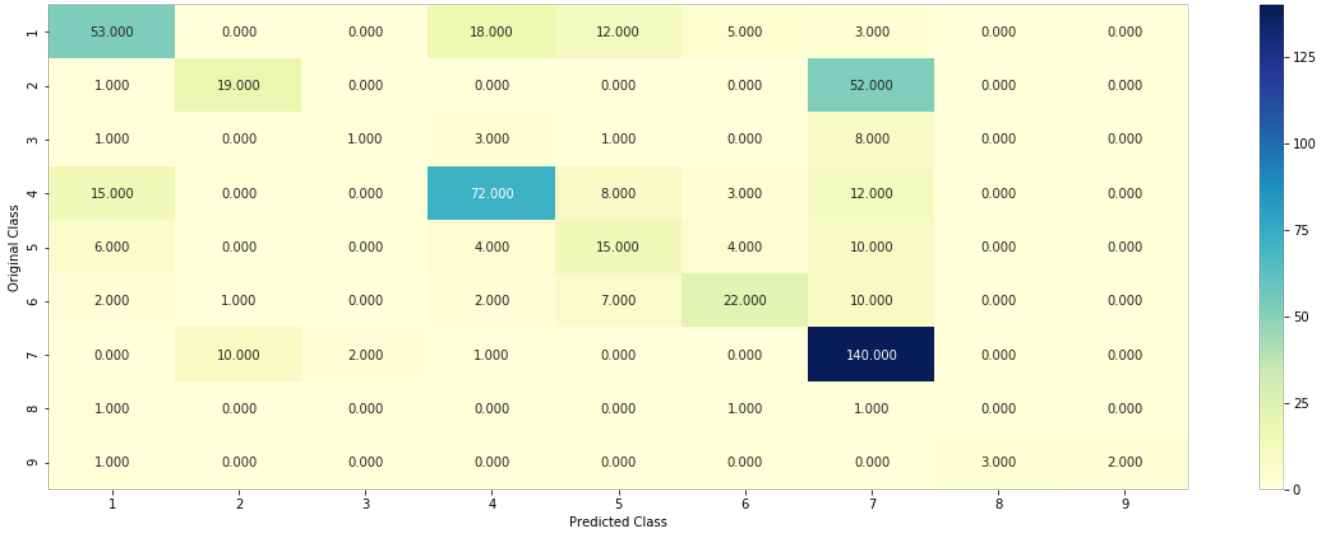

```
For values of best alpha =  0.0001 The train log loss is: 0.535873170364868
For values of best alpha =  0.0001 The cross validation log loss is: 1.0759275940084208
For values of best alpha =  0.0001 The test log loss is: 1.0979183347627088
```

Testing the model with best hyper paramters

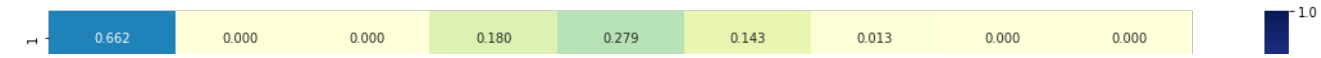
In [0]:

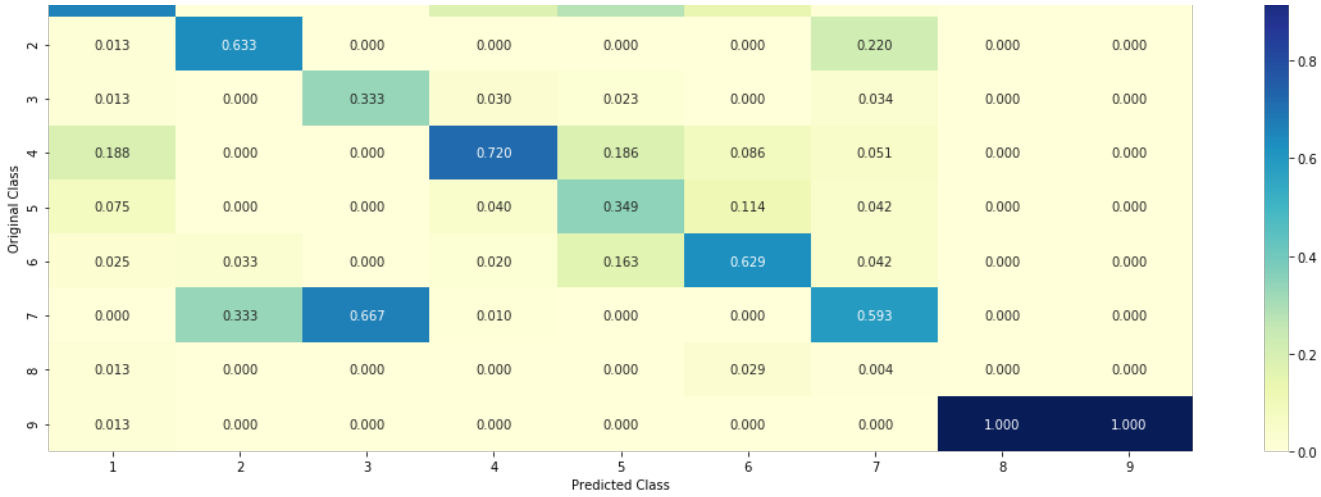
```
# read more about SGDClassifier() at http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDClassifier.html
# ------------------------------
# default parameters
# SGDClassifier(loss='hinge', penalty='l2', alpha=0.0001, l1_ratio=0.15, fit_intercept=True, max_iter=None, tol=None,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

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

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

```
Log loss : 1.0759275940084208
Number of mis-classified points : 0.39097744360902253
-------------------- Confusion matrix --------------------
```

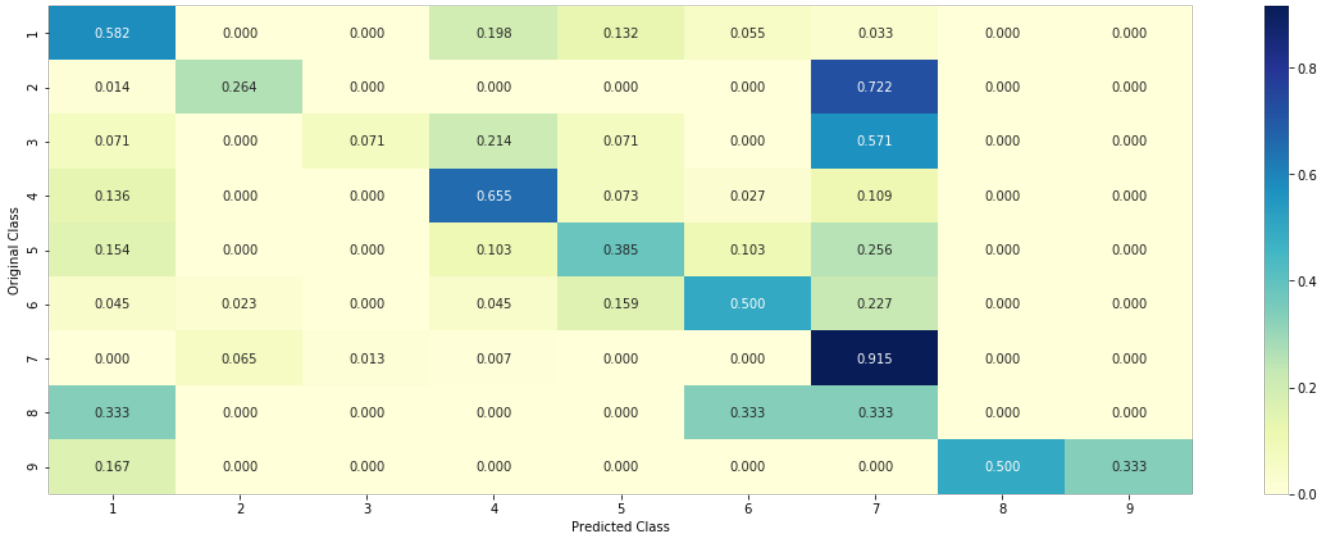



```
-------------------- Precision matrix (Columm Sum=1) --------------------
```





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



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 query point")
    print("-"*50)
    print("The features that are most importent of the ",predicted_cls[0]," class:")
    print (tabulate(tabulte_list, headers=["Index",'Feature name', 'Present or Not']))
```

Correctly Classified point

In [0]:

```
# from tabulate import tabulate
```

```
clf = SGDClassifier(class_weight='balanced', alpha=alpha[best_alpha], penalty='l2', loss='log', ran
dom_state=42)
clf.fit(train_x_tfidf,train_y)
test_point_index = 1
no_feature = 2000
predicted_cls = sig_clf.predict(test_x_tfidf[test_point_index])
print("Predicted Class :", predicted_cls[0])
print("Predicted Class Probabilities:",
np.round(sig_clf.predict_proba(test_x_tfidf[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],
test_df['TEXT'].iloc[test_point_index],test_df['Gene'].iloc[test_point_index],test_df['Variation']
.iloc[test_point_index], no_feature)
```

```
Predicted Class : 1
Predicted Class Probabilities: [[0.689  0.0576 0.0181 0.1138 0.0282 0.0393 0.0452 0.0043 0.0046]]
Actual Class : 1
--------------------------------------------------
346 Text feature [100] present in test data point [True]
Out of the top  2000  features  1 are present in query point
```

Incorrectly Classified point

In [0]:

```
test_point_index = 5
no_feature = 1000
predicted_cls = sig_clf.predict(test_x_tfidf[test_point_index])
print("Predicted Class :", predicted_cls[0])
print("Predicted Class Probabilities:",
np.round(sig_clf.predict_proba(test_x_tfidf[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],
test_df['TEXT'].iloc[test_point_index],test_df['Gene'].iloc[test_point_index],test_df['Variation']
.iloc[test_point_index], no_feature)
```

```
Predicted Class : 4
Predicted Class Probabilities: [[0.2231 0.0986 0.0335 0.3983 0.0621 0.0653 0.1062 0.0065 0.0064]]
Actual Class : 1
--------------------------------------------------
Out of the top  1000  features  0 are present in query point
```

5.2.3 Linear Support Vector Machines

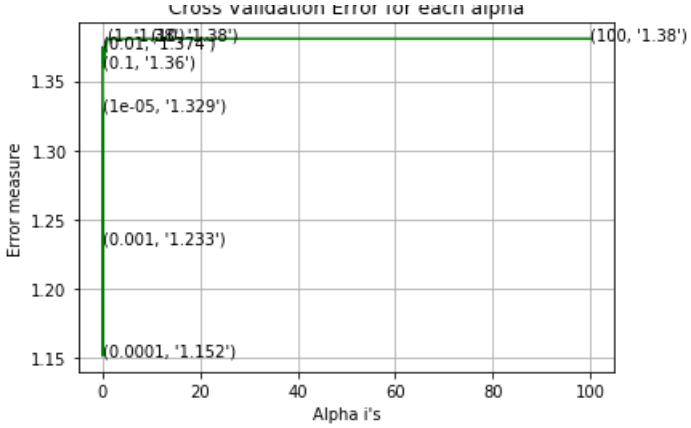
Hyper paramter tuning

In [0]:

```
# 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, t
ol=0.001,
# cache_size=200, class_weight=None, verbose=False, max_iter=-1, decision_function_shape='ovr', ra
ndom_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, method='sigmoid', cv=3)
#
# some of the methods of CalibratedClassifierCV()
# fit(X, y[, sample_weight]) Fit the calibrated model
# get_params([deep]) Get parameters for this estimator.
# predict(X) Predict the target of new samples.
# predict_proba(X) Posterior probabilities of classification
#-------------------------------------
# video link:
#-------------------------------------

alpha = [10 ** x for x in range(-5, 3)]
cv_log_error_array = []
for i in alpha:
    print("for C =", i)
    clf = SGDClassifier( class_weight='balanced', alpha=i, penalty='l2', loss='hinge', random_state
=42)
    clf.fit(train_x_tfidf, train_y)
    sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
    sig_clf.fit(train_x_tfidf, train_y)
    sig_clf_probs = sig_clf.predict_proba(cv_x_tfidf)
    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 = SVC(C=i,kernel='linear',probability=True, class_weight='balanced')
clf = SGDClassifier(class_weight='balanced', alpha=alpha[best_alpha], penalty='l2', loss='hinge', r
andom_state=42)
clf.fit(train_x_tfidf, train_y)
sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
sig_clf.fit(train_x_tfidf, train_y)

predict_y = sig_clf.predict_proba(train_x_tfidf)
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_tfidf)
print('For values of best alpha = ', alpha[best_alpha], "The cross validation log loss is:",log_lo
ss(y_cv, predict_y, labels=clf.classes_, eps=1e-15))
predict_y = sig_clf.predict_proba(test_x_tfidf)
print('For values of best alpha = ', alpha[best_alpha], "The test log loss is:",log_loss(y_test, p
redict_y, labels=clf.classes_, eps=1e-15))
```

```
for C = 1e-05
Log Loss : 1.3290999329449484
for C = 0.0001
Log Loss : 1.1518143108686003
for C = 0.001
Log Loss : 1.2325370533360203
for C = 0.01
Log Loss : 1.3744522313132166
for C = 0.1
Log Loss : 1.360099266650802
for C = 1
Log Loss : 1.380189642331655
for C = 10
Log Loss : 1.3803970364909934
for C = 100
Log Loss : 1.380397051055484
```



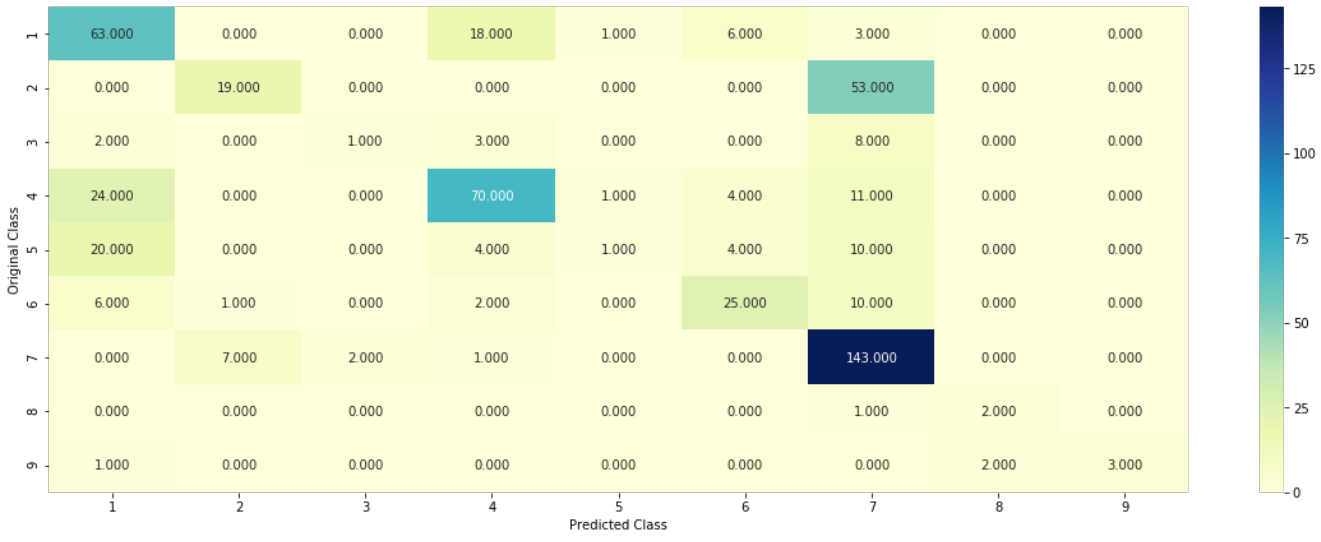

```
For values of best alpha =  0.0001 The train log loss is: 0.48910369321372754
For values of best alpha =  0.0001 The cross validation log loss is: 1.1518143108686003
For values of best alpha =  0.0001 The test log loss is: 1.1605020904239227
```

Testing model with best hyper parameters

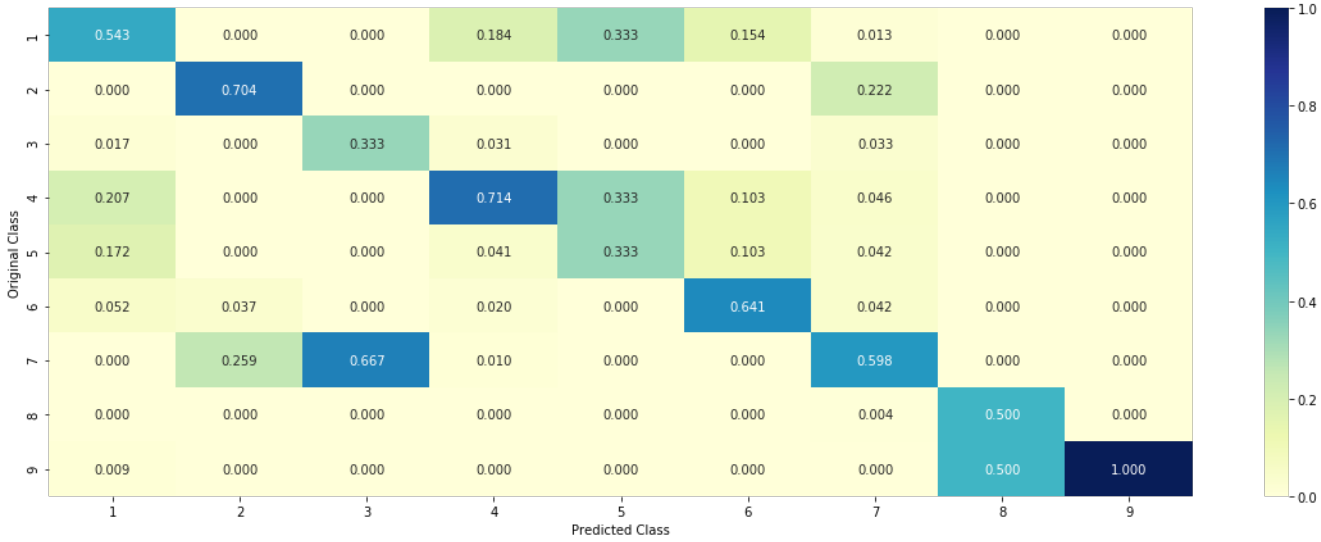
In [0]:

```
# 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, t
ol=0.001,
# cache_size=200, class_weight=None, verbose=False, max_iter=-1, decision_function_shape='ovr', ra
ndom_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/
# --------------------------------


# 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_tfidf, train_y,cv_x_tfidf,cv_y, clf)
```

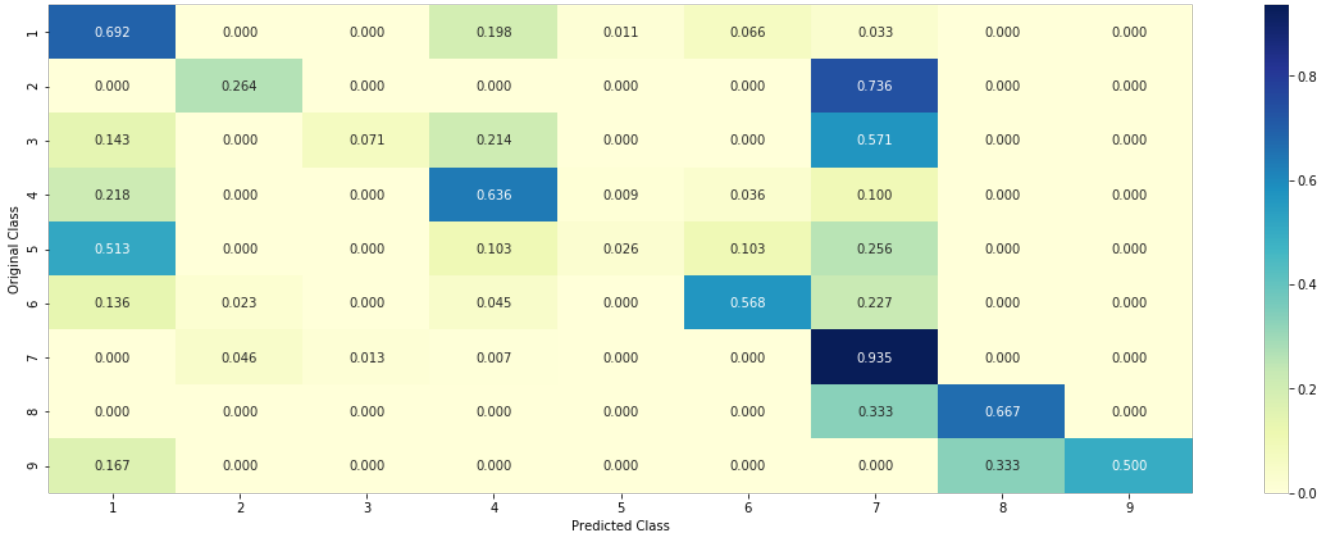
```
Log loss : 1.1518143108686003
Number of mis-classified points : 0.38533834586466165
-------------------- Confusion matrix --------------------
```



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




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



Feature Importance

For Correctly classified point

In [0]:

```
clf = SGDClassifier(alpha=alpha[best_alpha], penalty='l2', loss='hinge', random_state=42)
clf.fit(train_x_tfidf,train_y)
test_point_index = 1
# test_point_index = 100
no_feature = 500
predicted_cls = sig_clf.predict(test_x_tfidf[test_point_index])
print("Predicted Class :", predicted_cls[0])
print("Predicted Class Probabilities:",
np.round(sig_clf.predict_proba(test_x_tfidf[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],
test_df['TEXT'].iloc[test_point_index],test_df['Gene'].iloc[test_point_index],test_df['Variation']
.iloc[test_point_index], no_feature)
```

```
Predicted Class : 1
Predicted Class Probabilities: [[0.6459 0.1689 0.0077 0.0191 0.0107 0.0105 0.1033 0.0116 0.0222]]
Actual Class : 1
```

```
--------------------------------------------------
292 Text feature [000] present in test data point [True]
342 Text feature [004] present in test data point [True]
Out of the top  500  features  2 are present in query point
```

For Incorrectly classified point

In [0]:

```
test_point_index = 5
no_feature = 500
predicted_cls = sig_clf.predict(test_x_tfidf[test_point_index])
print("Predicted Class :", predicted_cls[0])
print("Predicted Class Probabilities:",
np.round(sig_clf.predict_proba(test_x_tfidf[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],
test_df['TEXT'].iloc[test_point_index],test_df['Gene'].iloc[test_point_index],test_df['Variation']
.iloc[test_point_index], no_feature)
```

```
Predicted Class : 4
Predicted Class Probabilities: [[0.194  0.0784 0.0194 0.5104 0.0539 0.0443 0.0883 0.005  0.0063]]
Actual Class : 1
--------------------------------------------------
Out of the top  500  features  0 are present in query point
```

5.2.4 Random Forest Classifier

Hyper paramter tuning

In [0]:

```
# --------------------------------
# default parameters
# sklearn.ensemble.RandomForestClassifier(n_estimators=10, criterion='gini', max_depth=None, min_s
amples_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-fores
t-and-their-construction-2/
# --------------------------------


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

```
# 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 , class_weight='balanced')
        clf.fit(train_x_tfidf, train_y)
        sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
        sig_clf.fit(train_x_tfidf, train_y)
        sig_clf_probs = sig_clf.predict_proba(cv_x_tfidf)
        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))


best_alpha = np.argmin(cv_log_error_array)
clf = RandomForestClassifier(n_estimators=alpha[int(best_alpha/2)], criterion='gini', max_depth=max
_depth[int(best_alpha%2)], random_state=42, n_jobs=-1)
clf.fit(train_x_tfidf, train_y)
sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
sig_clf.fit(train_x_tfidf, train_y)

predict_y = sig_clf.predict_proba(train_x_tfidf)
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_tfidf)
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_tfidf)
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 : 1.6140498101908118
for n_estimators = 100 and max depth =  10
Log Loss : 1.4421281473153638
for n_estimators = 200 and max depth =  5
Log Loss : 1.5897432973250707
for n_estimators = 200 and max depth =  10
Log Loss : 1.4135486429910085
for n_estimators = 500 and max depth =  5
Log Loss : 1.5738013983539314
for n_estimators = 500 and max depth =  10
Log Loss : 1.3845644085231155
for n_estimators = 1000 and max depth =  5
Log Loss : 1.5441488298720099
for n_estimators = 1000 and max depth =  10
Log Loss : 1.3742538138470832
for n_estimators = 2000 and max depth =  5
Log Loss : 1.5222705926762523
for n_estimators = 2000 and max depth =  10
Log Loss : 1.3717699980005245
For values of best estimator =  2000 The train log loss is: 1.0403450850922633
For values of best estimator =  2000 The cross validation log loss is: 1.262973166217094
For values of best estimator =  2000 The test log loss is: 1.269331292292979
```

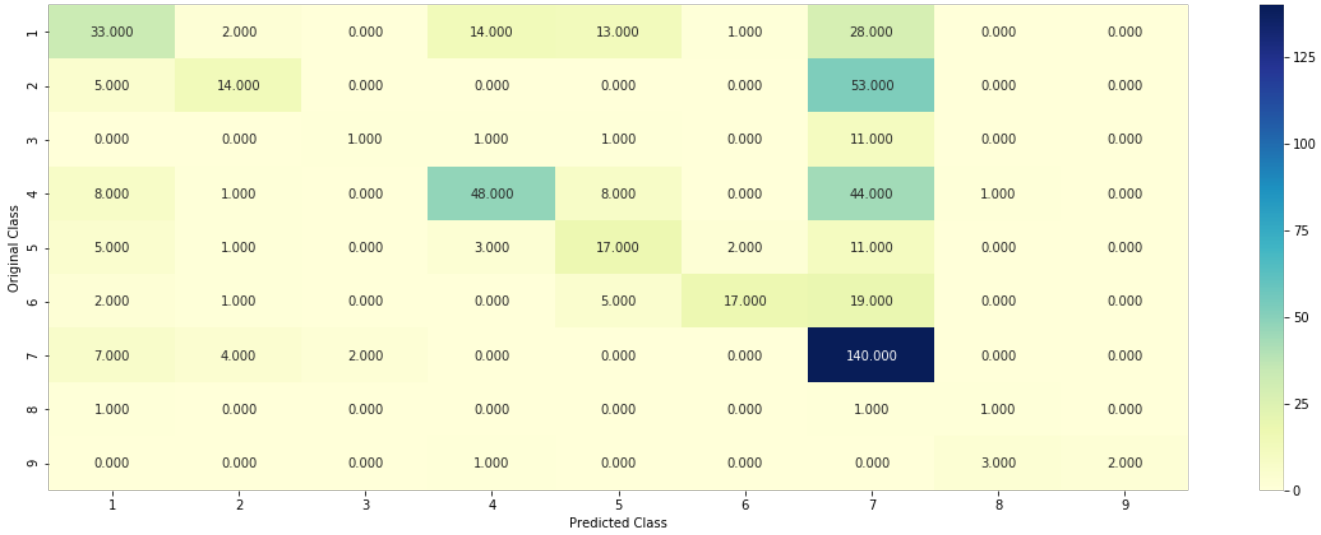
Testing model with best hyper parameters

In [0]:

```
# ---------------------------------
# default parameters
# sklearn.ensemble.RandomForestClassifier(n_estimators=10, criterion='gini', max_depth=None, min_s
amples_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()
```

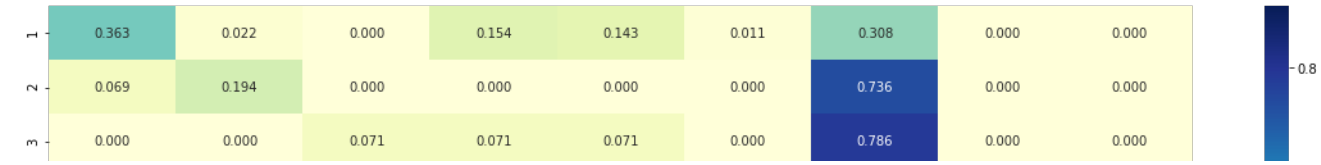
```
# fit(X, y, [sample_weight]) Fit the SVM model according to the given training data.
# predict(X) Perform classification on samples in X.
# predict_proba (X) Perform classification on samples in X.

# 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/
# --------------------------------

clf = RandomForestClassifier(n_estimators=alpha[int(best_alpha/2)], criterion='gini', max_depth=max_depth[int(best_alpha%2)], random_state=42, n_jobs=-1 , class_weight = 'balanced')
predict_and_plot_confusion_matrix(train_x_tfidf, train_y,cv_x_tfidf,cv_y, clf)
```

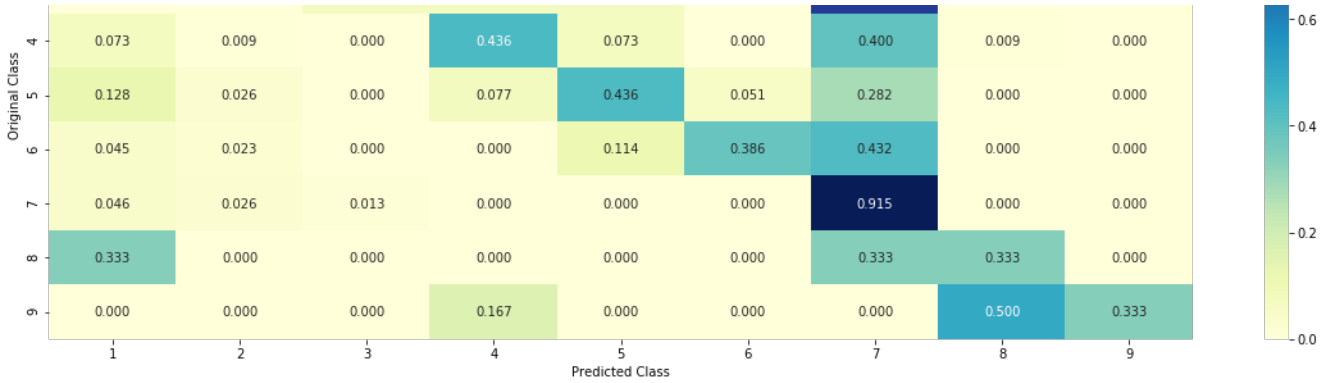
```
Log loss : 1.371769998000525
Number of mis-classified points : 0.4868421052631579
-------------------- Confusion matrix --------------------
```




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


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





Feature Importance

Incorrectly Classified point

In [0]:

```
# test_point_index = 10
clf = RandomForestClassifier(n_estimators=alpha[int(best_alpha/2)], criterion='gini', max_depth=max
_depth[int(best_alpha%2)], random_state=42, n_jobs=-1 , class_weight = 'balanced')
clf.fit(train_x_tfidf, train_y)
sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
sig_clf.fit(train_x_tfidf, train_y)

test_point_index = 5
no_feature = 100
predicted_cls = sig_clf.predict(test_x_tfidf[test_point_index])
print("Predicted Class :", predicted_cls[0])
print("Predicted Class Probabilities:",
np.round(sig_clf.predict_proba(test_x_tfidf[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], test_df['TEXT'].iloc[test_point_index],test_df['Gene'].
iloc[test_point_index],test_df['Variation'].iloc[test_point_index], no_feature)
```

```
Predicted Class : 7
Predicted Class Probabilities: [[0.141  0.1389 0.0273 0.2106 0.0764 0.0702 0.3248 0.0048 0.006 ]]
Actual Class : 1
--------------------------------------------------
40 Text feature [12] present in test data point [True]
Out of the top  100  features  1 are present in query point
```

Correctly Classified point

In [0]:

```
test_point_index = 5
no_feature = 100
predicted_cls = sig_clf.predict(test_x_tfidf[test_point_index])
print("Predicted Class :", predicted_cls[0])
print("Predicted Class Probabilities:",
np.round(sig_clf.predict_proba(test_x_tfidf[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], test_df['TEXT'].iloc[test_point_index],test_df['Gene'].
iloc[test_point_index],test_df['Variation'].iloc[test_point_index], no_feature)
```

```
Predicted Class : 2
Predicted Class Probabilities: [[0.0503 0.5546 0.0365 0.0353 0.034  0.0445 0.2368 0.0035 0.0045]]
Actual Class : 2
--------------------------------------------------
59 Text feature [116] present in test data point [True]
Out of the top  100  features  1 are present in query point
```

5.2.5 Stack the models

Testing with hyper parameter tuning

In [0]:

```
# 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_i
ter=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-in
tuition-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, t
ol=0.001,
# cache_size=200, class_weight=None, verbose=False, max_iter=-1, decision_function_shape='ovr', ra
ndom_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_s
amples_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-fores
t-and-their-construction-2/
# ---------------------------------


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

```
clf2 = SGDClassifier(alpha=1, penalty='l2', loss='hinge', class_weight='balanced', random_state=0)
clf2.fit(train_x_tfidf, train_y)
sig_clf2 = CalibratedClassifierCV(clf2, method="sigmoid")


clf3 = MultinomialNB(alpha=0.001)
clf3.fit(train_x_tfidf, train_y)
sig_clf3 = CalibratedClassifierCV(clf3, method="sigmoid")

sig_clf1.fit(train_x_tfidf, train_y)
print("Logistic Regression :  Log Loss: %0.2f" % (log_loss(cv_y, sig_clf1.predict_proba(cv_x_tfidf)
)))
sig_clf2.fit(train_x_tfidf, train_y)
print("Support vector machines : Log Loss: %0.2f" % (log_loss(cv_y,
sig_clf2.predict_proba(cv_x_tfidf))))
sig_clf3.fit(train_x_tfidf, train_y)
print("Naive Bayes : Log Loss: %0.2f" % (log_loss(cv_y, sig_clf3.predict_proba(cv_x_tfidf))))
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_p
robas=True)
    sclf.fit(train_x_tfidf, train_y)
    print("Stacking Classifer : for the value of alpha: %f Log Loss: %0.3f" % (i, log_loss(cv_y, sc
lf.predict_proba(cv_x_tfidf))))
    log_error =log_loss(cv_y, sclf.predict_proba(cv_x_tfidf))
    if best_alpha > log_error:
        best_alpha = log_error
```

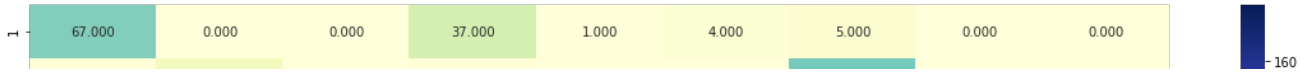
```
Logistic Regression :  Log Loss: 1.12
Support vector machines : Log Loss: 1.38
Naive Bayes : Log Loss: 1.24
--------------------------------------------------
Stacking Classifer : for the value of alpha: 0.000100 Log Loss: 2.179
Stacking Classifer : for the value of alpha: 0.001000 Log Loss: 2.044
Stacking Classifer : for the value of alpha: 0.010000 Log Loss: 1.561
Stacking Classifer : for the value of alpha: 0.100000 Log Loss: 1.198
Stacking Classifer : for the value of alpha: 1.000000 Log Loss: 1.331
Stacking Classifer : for the value of alpha: 10.000000 Log Loss: 1.761
```

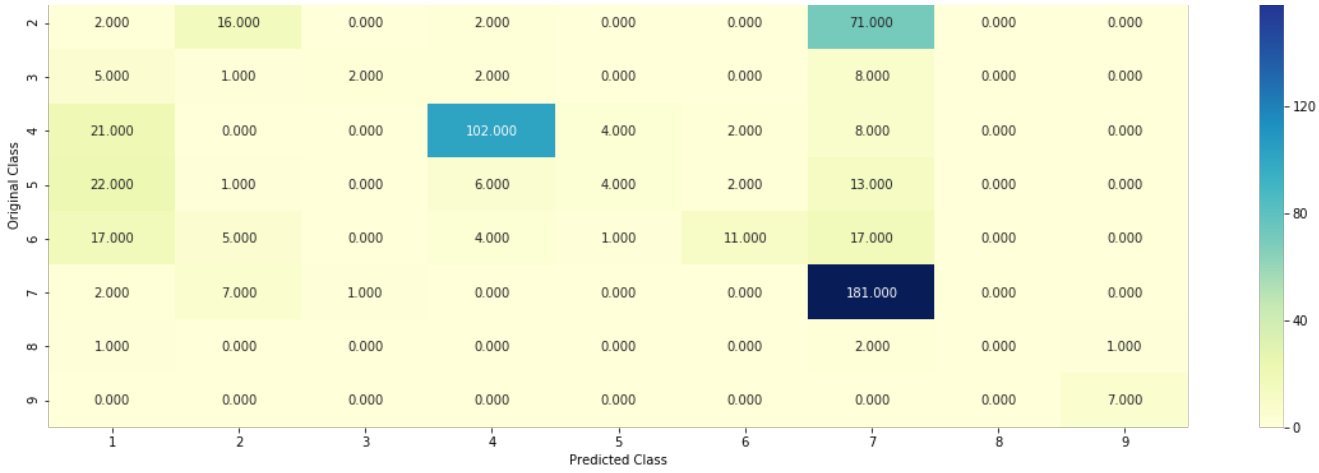
Testing the model with the best hyper parameters

In [0]:

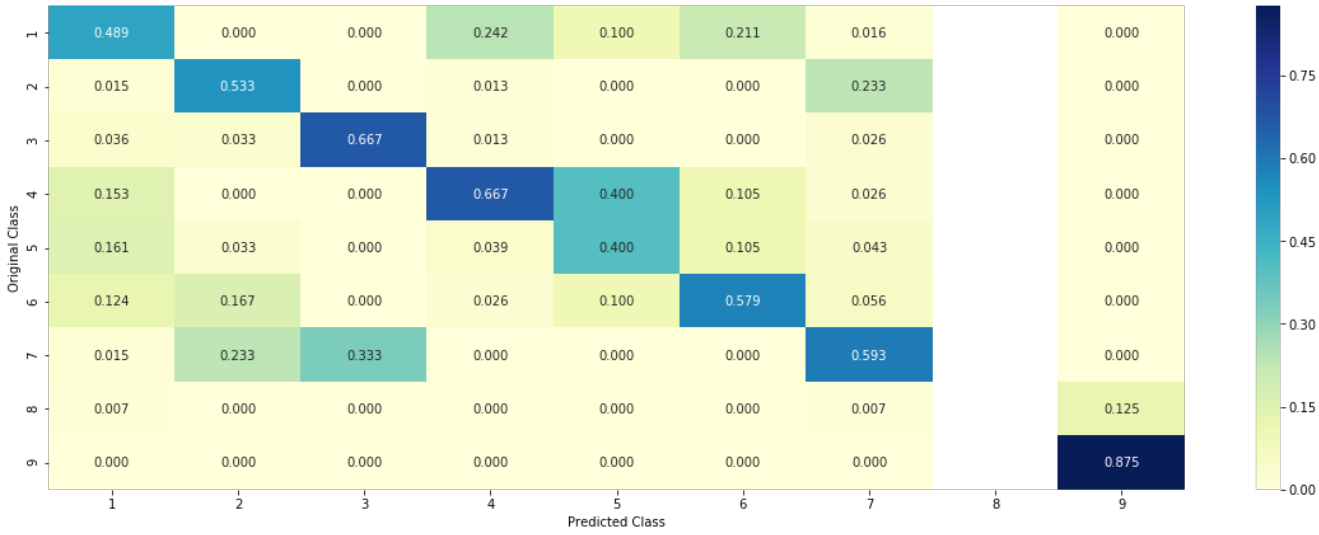
```
lr = LogisticRegression(C=0.1)
sclf = StackingClassifier(classifiers=[sig_clf1, sig_clf2, sig_clf3], meta_classifier=lr, use_proba
s=True)
sclf.fit(train_x_tfidf, train_y)

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

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

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

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

```
Log loss (train) on the stacking classifier : 0.5489928209703522
Log loss (CV) on the stacking classifier : 1.197531350909727
Log loss (test) on the stacking classifier : 1.2237383803446504
Number of missclassified point : 0.41353383458646614
-------------------- Confusion matrix --------------------
```

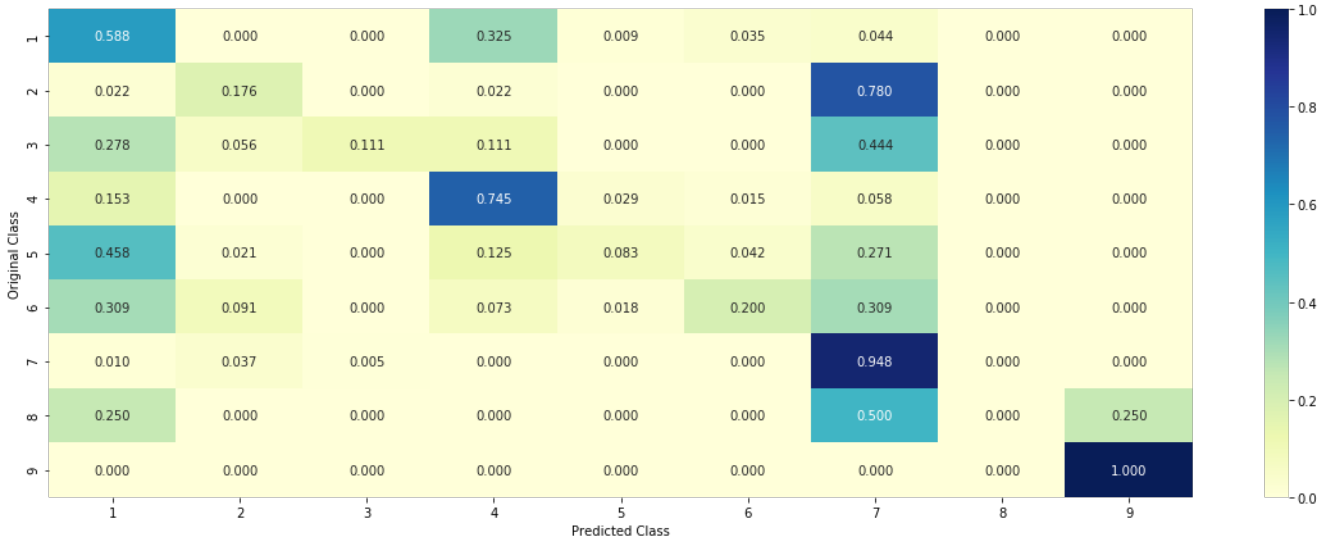




---------- Precision matrix (Columm Sum=1) ----------



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



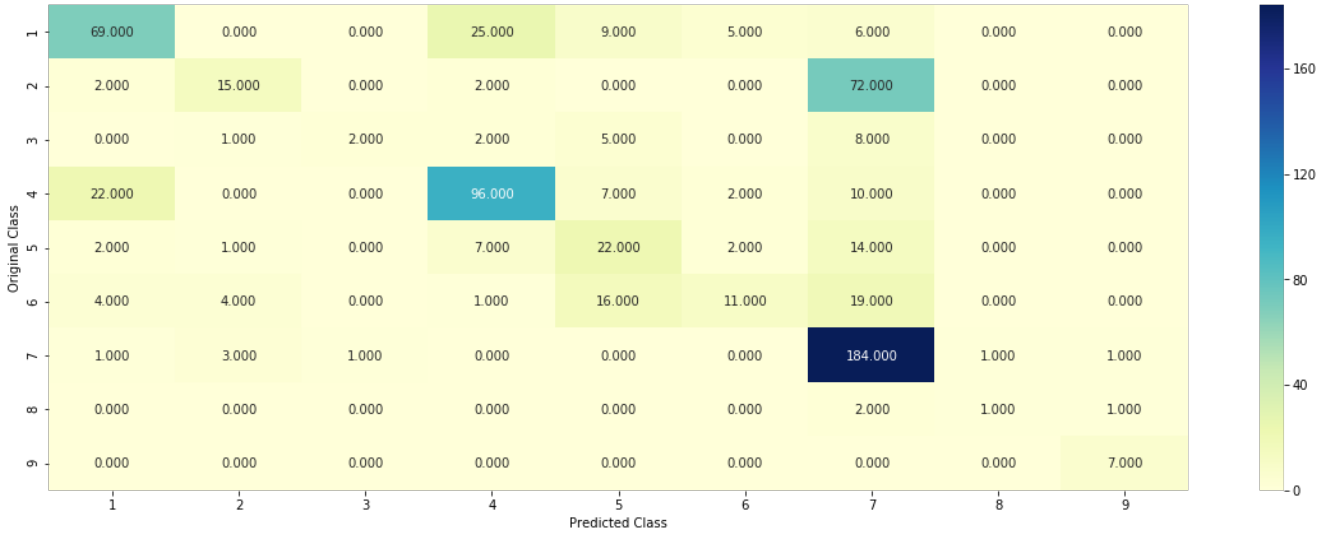
5.2.5.2 Maximum Voting classifier

In [0]:

```
#Refer:http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.VotingClassifier.html
from sklearn.ensemble import VotingClassifier
vclf = VotingClassifier(estimators=[('lr', sig_clf1), ('svc', sig_clf2), ('rf', sig_clf3)], voting=
'soft')
```

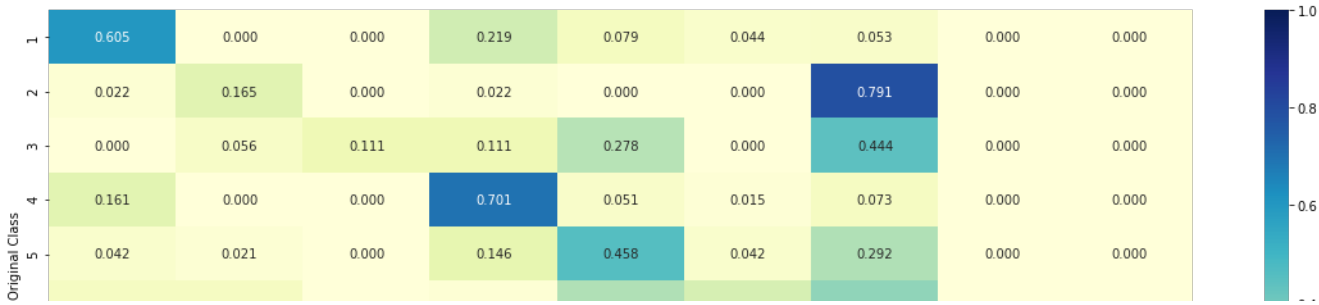
```
vclf.fit(train_x_tfidf, train_y)
print("Log loss (train) on the VotingClassifier :", log_loss(train_y,
vclf.predict_proba(train_x_tfidf)))
print("Log loss (CV) on the VotingClassifier :", log_loss(cv_y, vclf.predict_proba(cv_x_tfidf)))
print("Log loss (test) on the VotingClassifier :", log_loss(test_y,
vclf.predict_proba(test_x_tfidf)))
print("Number of missclassified point :", np.count_nonzero((vclf.predict(test_x_tfidf)- test_y))/t
est_y.shape[0])
plot_confusion_matrix(test_y=test_y, predict_y=vclf.predict(test_x_tfidf))
```

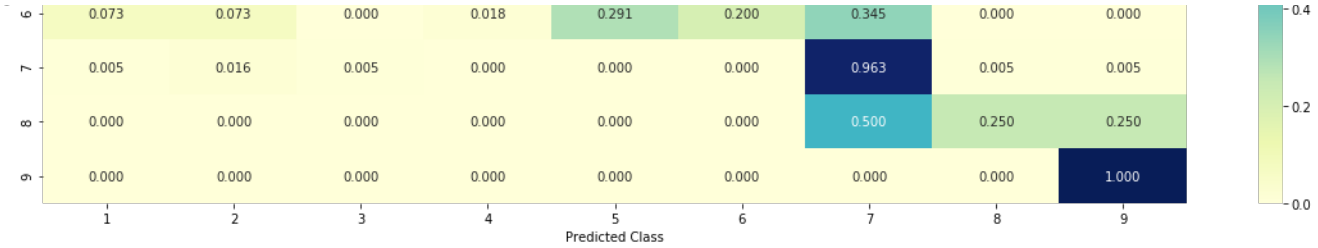
```
Log loss (train) on the VotingClassifier : 0.8134672955197163
Log loss (CV) on the VotingClassifier : 1.193380311803099
Log loss (test) on the VotingClassifier : 1.2099736492629896
Number of missclassified point : 0.3879699248120301
-------------------- Confusion matrix --------------------
```




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


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





5.3 Using both unigrams and bigrams to make the text to vectors and Logistic regression on top of it

In [0]:

```
vectorizer = CountVectorizer(ngram_range=(1,2) , min_df=3)
train_text_feature_ngram = vectorizer.fit_transform(train_df['TEXT'])
# getting all the feature names (words)
train_text_features= vectorizer.get_feature_names()

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

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


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

```
Total number of unique words in train data : 761668
```

In [0]:

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

# dict_list[i] is build on i'th  class text data
# total_dict is buid on whole training text data
total_dict = extract_dictionary_paddle(train_df)


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]:

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

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

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

Stacking the vectors together

In [0]:

```
train_x_ngram = hstack((train_gene_var_onehotCoding, train_text_feature_ngram)).tocsr()

test_x_ngram = hstack((test_gene_var_onehotCoding, test_text_feature_ngram)).tocsr()

cv_x_ngram = hstack((cv_gene_var_onehotCoding, cv_text_feature_ngram)).tocsr()
```

In [0]:

```
print("TFIDF features :")
print("(number of data points * number of features) in train data = ", train_x_ngram.shape)
print("(number of data points * number of features) in test data = ", test_x_ngram.shape)
print("(number of data points * number of features) in cross validation data =", cv_x_ngram.shape)
```

```
TFIDF features :
(number of data points * number of features) in train data =  (2124, 763855)
(number of data points * number of features) in test data =  (665, 763855)
(number of data points * number of features) in cross validation data = (532, 763855)
```

5.3.1. Logistic Regression

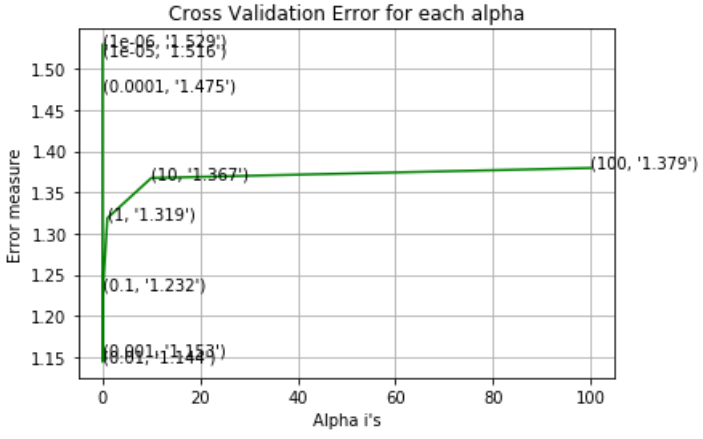
Hyper paramter tuning

In [0]:

```
# 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_i
ter=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-in
tuition-1/
#-------------------------------


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

alpha = [10 ** x for x in range(-6, 3)]
cv_log_error_array = []
for i in alpha:
    print("for alpha =", i)
    clf = SGDClassifier(class_weight='balanced', alpha=i, penalty='l2', loss='log', random_state=42
)
    clf.fit(train_x_ngram, train_y)
    sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
```

```
    sig_clf.fit(train_x_ngram, train_y)
    sig_clf_probs = sig_clf.predict_proba(cv_x_ngram)
    cv_log_error_array.append(log_loss(cv_y, sig_clf_probs, labels=clf.classes_, eps=1e-15))
    # to avoid rounding error while multiplying probabilites we use log-probability estimates
    print("Log Loss :",log_loss(cv_y, sig_clf_probs))

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


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

predict_y = sig_clf.predict_proba(train_x_ngram)
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_ngram)
print('For values of best alpha = ', alpha[best_alpha], "The cross validation log loss is:",log_lo
ss(y_cv, predict_y, labels=clf.classes_, eps=1e-15))
predict_y = sig_clf.predict_proba(test_x_ngram)
print('For values of best alpha = ', alpha[best_alpha], "The test log loss is:",log_loss(y_test, p
redict_y, labels=clf.classes_, eps=1e-15))
```

```
for alpha = 1e-06
Log Loss : 1.5290985391163852
for alpha = 1e-05
Log Loss : 1.5159741010523877
for alpha = 0.0001
Log Loss : 1.4745435435436645
for alpha = 0.001
Log Loss : 1.1532457714319013
for alpha = 0.01
Log Loss : 1.1440925447451045
for alpha = 0.1
Log Loss : 1.2315346318312645
for alpha = 1
Log Loss : 1.3185703422726476
for alpha = 10
Log Loss : 1.3673250973745859
for alpha = 100
Log Loss : 1.3794453951592283
```



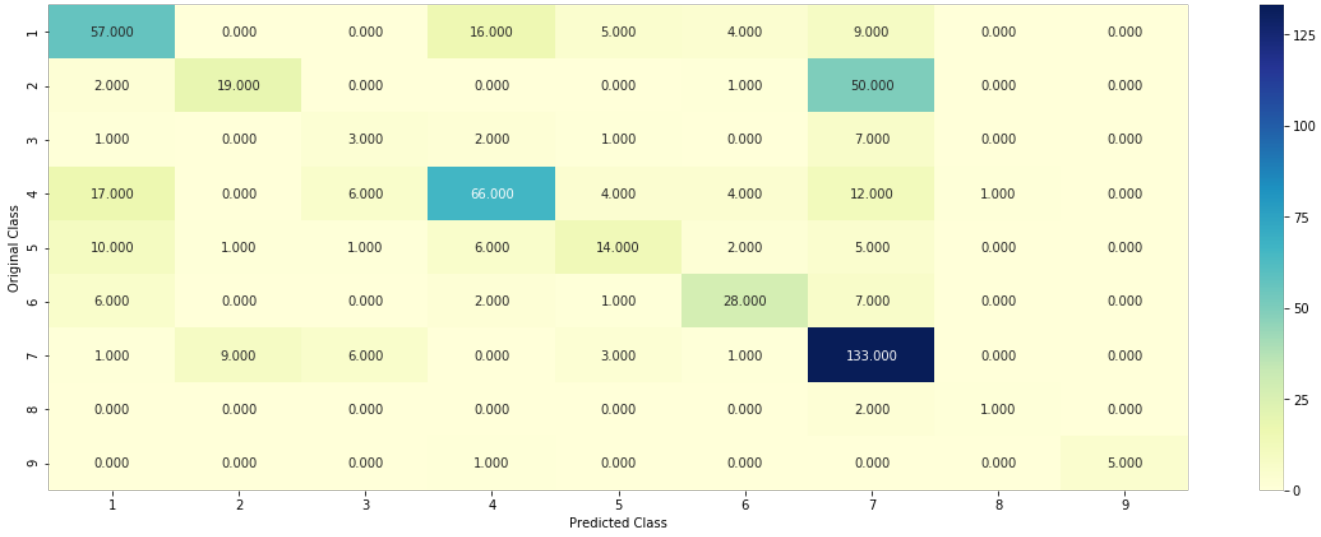

```
For values of best alpha =  0.01 The train log loss is: 0.6914717680735984
For values of best alpha =  0.01 The cross validation log loss is: 1.1440925447451045
For values of best alpha =  0.01 The test log loss is: 1.2154442466777582
```

Testing the model with best hyper paramters

In [0]:

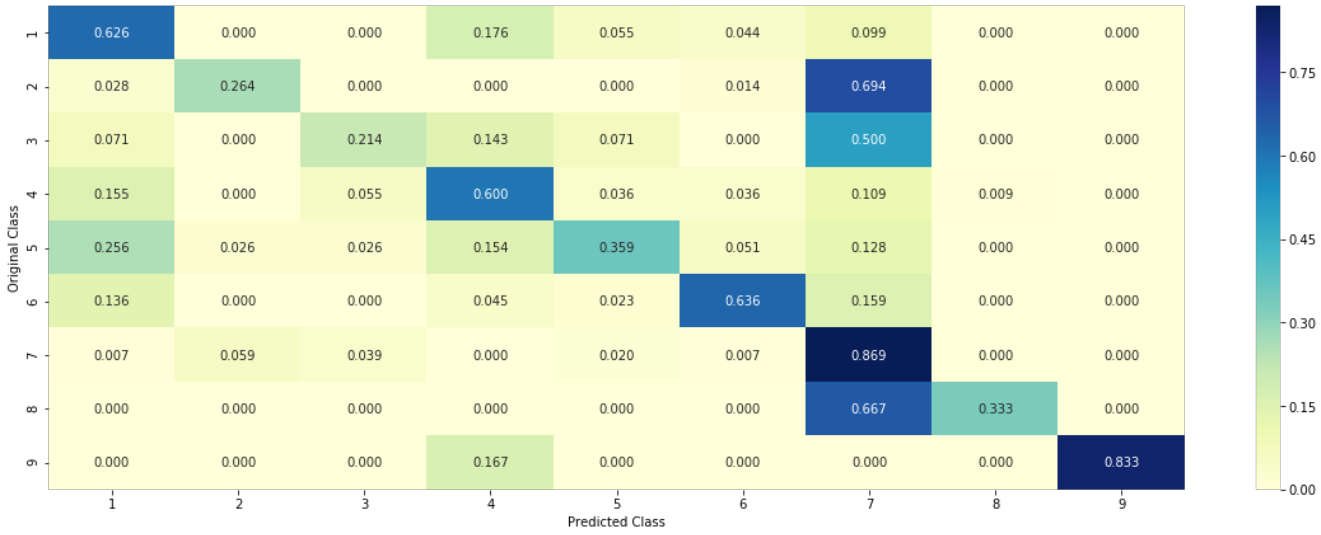
```
# read more about SGDClassifier() at http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDClassifier.html
# ------------------------------
# default parameters
# SGDClassifier(loss='hinge', penalty='l2', alpha=0.0001, l1_ratio=0.15, fit_intercept=True, max_iter=None, tol=None,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

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

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

```
Log loss : 1.1440925447451045
Number of mis-classified points : 0.38721804511278196
-------------------- Confusion matrix --------------------
```




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

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



5.4 Applying Feature Engineering to reduce test loss

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)
)
```

In [0]:

```
vectorizer = TfidfVectorizer(ngram_range=(1,4) , min_df=3 , max_features=6000)
train_text_feature_ngram = vectorizer.fit_transform(train_df['TEXT'])
# getting all the feature names (words)
train_text_features= vectorizer.get_feature_names()

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

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


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

```
Total number of unique words in train data : 6000
```

In [0]:

```
test_text_feature_ngram = vectorizer.transform(test_df['TEXT'])
cv_text_feature_ngram = vectorizer.transform(cv_df['TEXT'])
```

In [0]:

```
from sklearn.feature_selection import SelectKBest
from sklearn.feature_selection import chi2
selection = SelectKBest(chi2, k=2000).fit(train_text_feature_ngram, y_train)

train_text_feature_ngram = selection.transform(train_text_feature_ngram)
test_text_feature_ngram = selection.transform(test_text_feature_ngram)
cv_text_feature_ngram = selection.transform(cv_text_feature_ngram)

train_text_feature_ngram = normalize(train_text_feature_ngram, axis=0)
```

```
test_text_feature_ngram = normalize(test_text_feature_ngram, axis=0)
cv_text_feature_ngram = normalize(cv_text_feature_ngram, axis=0)
```

In [0]:

```
train_y = np.array(list(train_df['Class']))
test_y = np.array(list(test_df['Class']))
cv_y = np.array(list(cv_df['Class']))
```

Stacking the vectors together

In [0]:

```
train_x_ngram = hstack((train_gene_var_onehotCoding, train_text_feature_ngram)).tocsr()

test_x_ngram = hstack((test_gene_var_onehotCoding, test_text_feature_ngram)).tocsr()

cv_x_ngram = hstack((cv_gene_var_onehotCoding, cv_text_feature_ngram)).tocsr()
```

In [0]:

```
print("TFIDF features :")
print("(number of data points * number of features) in train data = ", train_x_ngram.shape)
print("(number of data points * number of features) in test data = ", test_x_ngram.shape)
print("(number of data points * number of features) in cross validation data =", cv_x_ngram.shape)
```

```
TFIDF features :
(number of data points * number of features) in train data =  (2124, 4208)
(number of data points * number of features) in test data =  (665, 4208)
(number of data points * number of features) in cross validation data = (532, 4208)
```

5.4.1. Logistic Regression

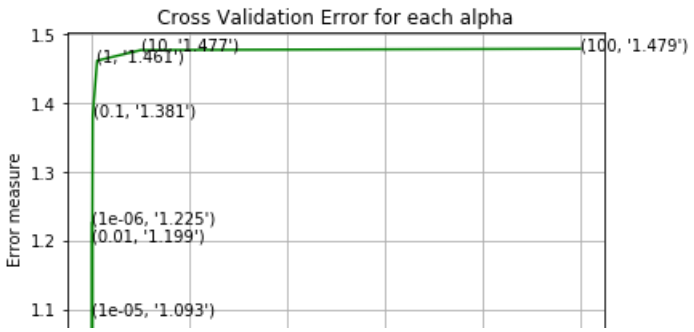
Hyper paramter tuning

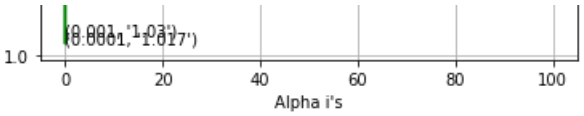
In [0]:

```
# 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_i
ter=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-in
tuition-1/
#-------------------------------


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

```
# 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_ngram, train_y)
    sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
    sig_clf.fit(train_x_ngram, train_y)
    sig_clf_probs = sig_clf.predict_proba(cv_x_ngram)
    cv_log_error_array.append(log_loss(cv_y, sig_clf_probs, labels=clf.classes_, eps=1e-15))
    # to avoid rounding error while multiplying probabilites we use log-probability estimates
    print("Log Loss :",log_loss(cv_y, sig_clf_probs))

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


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

predict_y = sig_clf.predict_proba(train_x_ngram)
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_ngram)
print('For values of best alpha = ', alpha[best_alpha], "The cross validation log loss is:",log_lo
ss(y_cv, predict_y, labels=clf.classes_, eps=1e-15))
predict_y = sig_clf.predict_proba(test_x_ngram)
print('For values of best alpha = ', alpha[best_alpha], "The test log loss is:",log_loss(y_test, p
redict_y, labels=clf.classes_, eps=1e-15))
```

```
for alpha = 1e-06
Log Loss : 1.2249793596543914
for alpha = 1e-05
Log Loss : 1.0934008116176148
for alpha = 0.0001
Log Loss : 1.0174559287083746
for alpha = 0.001
Log Loss : 1.0302828562667006
for alpha = 0.01
Log Loss : 1.198876956311648
for alpha = 0.1
Log Loss : 1.3810222298552894
for alpha = 1
Log Loss : 1.461222817036608
for alpha = 10
Log Loss : 1.4765592599628168
for alpha = 100
Log Loss : 1.4785082826204525
```





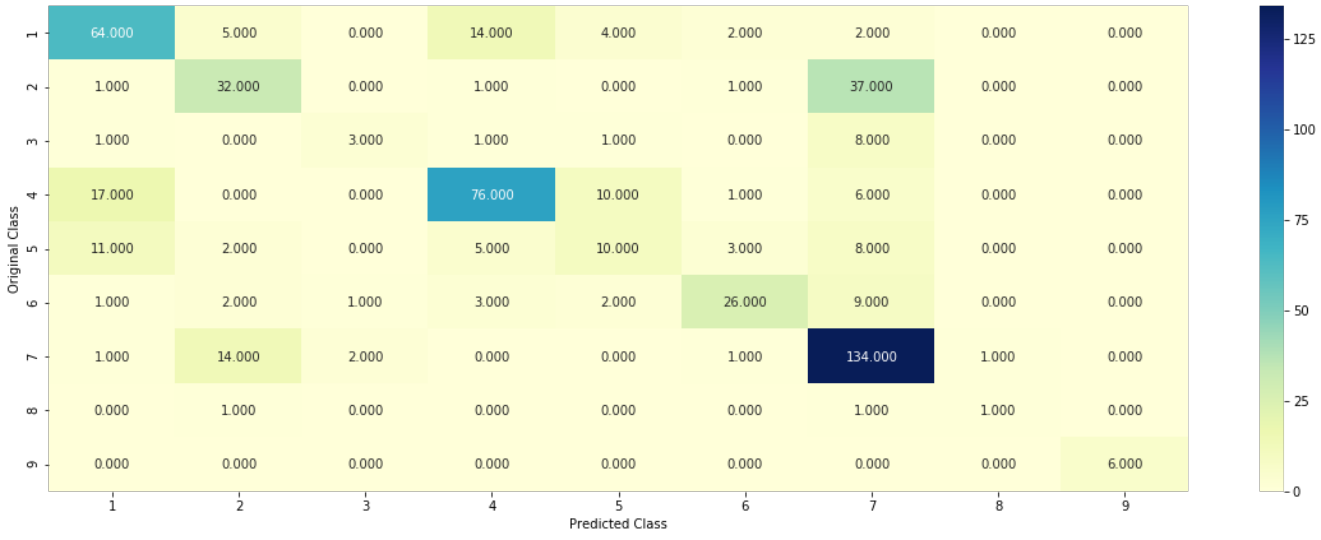

```
For values of best alpha =  0.0001 The train log loss is: 0.4119458061471544
For values of best alpha =  0.0001 The cross validation log loss is: 1.0174559287083746
For values of best alpha =  0.0001 The test log loss is: 0.9949116834424129
```

Testing the model with best hyper paramters

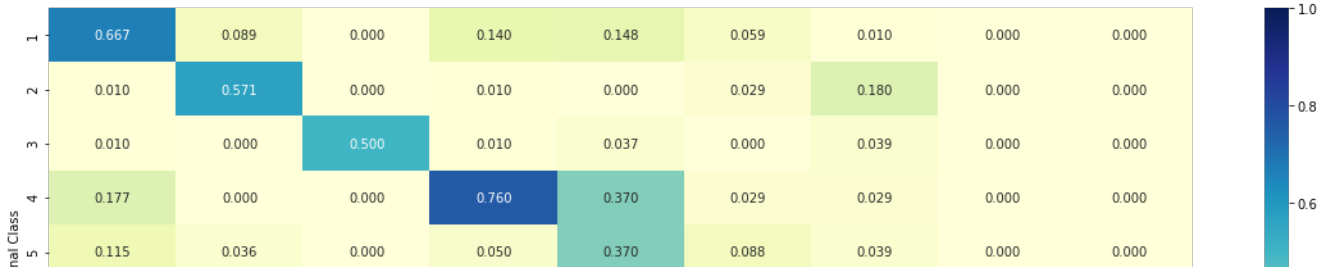
In [0]:

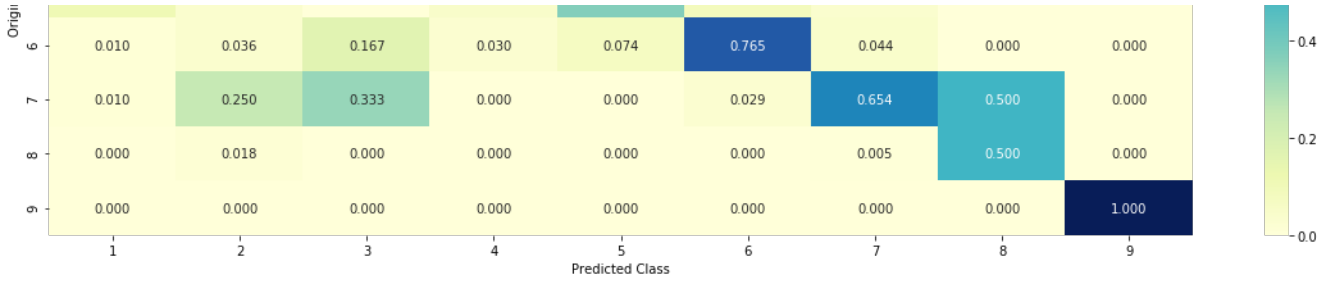
```
# read more about SGDClassifier() at http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDClassifier.html
# ------------------------------
# default parameters
# SGDClassifier(loss='hinge', penalty='l2', alpha=0.0001, l1_ratio=0.15, fit_intercept=True, max_iter=None, tol=None,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

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

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

```
Log loss : 1.0174559287083746
Number of mis-classified points : 0.3383458646616541
-------------------- Confusion matrix --------------------
```

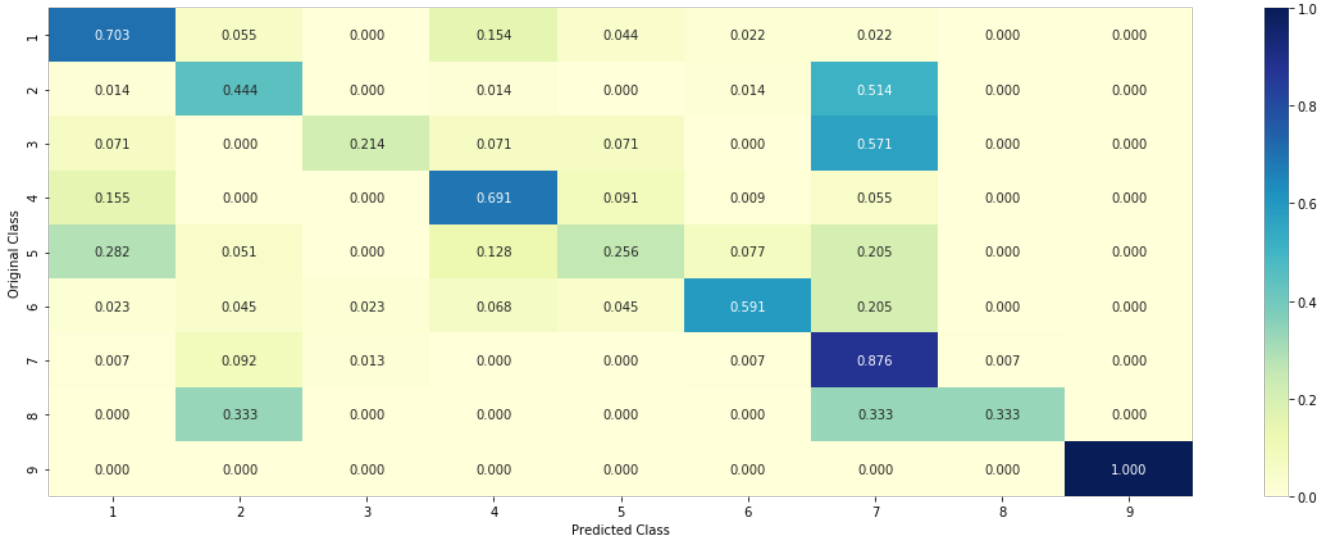



```
-------------------- Precision matrix (Column Sum=1) --------------------
```





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



6 Conclusion

In [0]:

```
from prettytable import PrettyTable
```

In [0]:

```
x = PrettyTable()
x.field_names = ["Vectorizer", "Algorithm", "Model", "Hyper Parameters", "Log-Loss", "Percentage of misclassified points"]
```

In [0]:

```
#4
x.add_row(["Bag of Words", "Naive Bayes", "Brute","alpha=0.1", 1.27, 44.36])
x.add_row(["Bag of Words", "KNN", "Brute","K=15", 1.09, 39.66])
x.add_row(["Bag of Words", "Logistic Regression + Class Balancing", "Brute","alpha=0.001", 1.10, 35.34])
x.add_row(["Bag of Words", "Logistic Regression", "Brute","alpha=0.001", 1.13, 37.27])
x.add_row(["Bag of Words", "Linear SVM", "Brute","alpha=0.01", 1.15, 39.47])
x.add_row(["Bag of Words", "Random Forest", "Brute","n_estimator=1000, max_depth=10", 1.16, 38.53])
x.add_row(["Response Coding", "Random Forest", "Brute","n_estimator=100, max_depth=5", 1.38, 51.8])
x.add_row(["Bag of Words", "Stacking Models", "Brute","LR[LR(alpha=0.01), SVM(alpha=1), MultinomialNB(alpha=0.001)]alpha=0.1", 1.15, 37.29])
x.add_row(["Bag of Words", "Maximum Voting", "Brute","LR(alpha=0.01), SVM(alpha=1), MultinomialNB(alpha=0.001)", 1.22, 38.19])

#5.1
x.add_row(["TFIDF", "Naive Bayes", "Brute","alpha=1000", 1.25, 41.91])
x.add_row(["TFIDF", "Logistic Regression + Class Balancing", "Brute","alpha=0.001", 1.11, 37.03])
x.add_row(["TFIDF", "Linear SVM", "Brute","alpha=0.001", 1.14, 35.33])
x.add_row(["TFIDF", "Random Forest", "Brute","n_estimator=2000, max_depth=10", 1.17, 40.41])
x.add_row(["TFIDF", "Stacking Models", "Brute","LR[LR(alpha=0.01), SVM(alpha=1), MultinomialNB(alpha=0.001)]alpha=0.1", 1.16, 38.94])
```

```
x.add_row(["TFIDF", "Maximum Voting", "Brute","LR(alpha=0.01), SVM(alpha=1),
MultinomialNB(alpha=0.001)", 1.16, 38.64])

#5.2
x.add_row(["TFIDF", "Naive Bayes", "SelectKBest top 1000","alpha=1", 1.17, 40.60])
x.add_row(["TFIDF", "Logistic Regression + Class Balancing", "SelectKBest top 1000","alpha=0.0001"
, 1.09, 39.09])
x.add_row(["TFIDF", "Linear SVM", "SelectKBest top 1000","alpha=0.0001", 1.16, 38.53])
x.add_row(["TFIDF", "Random Forest", "SelectKBest top 1000","n_estimator=2000, max_depth=10", 1.26
, 48.68])
x.add_row(["TFIDF", "Stacking Models", "SelectKBest top 1000","LR[LR(alpha=0.01), SVM(alpha=1), Mu
ltinomialNB(alpha=0.001)]alpha=0.1", 1.22, 41.35])
x.add_row(["TFIDF", "Maximum Voting", "SelectKBest top 1000","LR(alpha=0.01), SVM(alpha=1),
MultinomialNB(alpha=0.001)", 1.20, 38.79])

#5.3
x.add_row(["Bag of Words", "Logistic Regression + Class Balancing", "Unigram +Bigram","alpha=0.01"
, 1.21, 38.72])

#5.4
x.add_row(["Bag of Words", "Logistic Regression + Class Balancing", "(1,2,3,4) -
grams","alpha=0.0001", 0.99, 33.83])
```

In [0]:

```
print(x)
```

```
+-----------------+--------------------------------------------+----------------------+-----------------
-------------------------------------------------+----------+-------------------------------------+
|    Vectorizer   |                 Algorithm                  |        Model         |
Hyper Parameters                            | Log-Loss | Percentage of misclassified points |
+-----------------+--------------------------------------------+----------------------+-----------------
-------------------------------------------------+----------+-------------------------------------+
|   Bag of Words  |                Naive Bayes                 |        Brute         |
alpha=0.1                                   |   1.27   |               44.36                 |
|   Bag of Words  |                    KNN                     |        Brute         |
K=15                                        |   1.09   |              39.66                  |
|   Bag of Words  | Logistic Regression + Class Balancing      |        Brute         |
alpha=0.001                                 |   1.1    |               35.34                 |
|   Bag of Words  |          Logistic Regression               |        Brute         |
alpha=0.001                                 |   1.13   |               37.27                 |
|   Bag of Words  |                Linear SVM                  |        Brute         |
alpha=0.01                                  |   1.15   |               39.47                 |
|   Bag of Words  |             Random Forest                  |        Brute         |
n_estimator=1000, max_depth=10              |   1.16   |               38.53                 |
| Response Coding |             Random Forest                  |        Brute         |
n_estimator=100, max_depth=5                |   1.38   |               51.8                  |
|   Bag of Words  |            Stacking Models                 |        Brute         |
LR[LR(alpha=0.01), SVM(alpha=1), MultinomialNB(alpha=0.001)]alpha=0.1 |   1.15   |               37
.29                 |
|   Bag of Words  |             Maximum Voting                 |        Brute         |
LR(alpha=0.01), SVM(alpha=1), MultinomialNB(alpha=0.001)       |   1.22   |               38.19
|
|      TFIDF      |                Naive Bayes                 |        Brute         |
alpha=1000                                  |   1.25   |               41.91                 |
|      TFIDF      | Logistic Regression + Class Balancing      |        Brute         |
alpha=0.001                                 |   1.11   |               37.03                 |
|      TFIDF      |                Linear SVM                  |        Brute         |
alpha=0.001                                 |   1.14   |               35.33                 |
|      TFIDF      |             Random Forest                  |        Brute         |
n_estimator=2000, max_depth=10              |   1.17   |               40.41                 |
|      TFIDF      |            Stacking Models                 |        Brute         |
LR[LR(alpha=0.01), SVM(alpha=1), MultinomialNB(alpha=0.001)]alpha=0.1 |   1.16   |               38
.94                 |
|      TFIDF      |             Maximum Voting                 |        Brute         |        LR(alpha=
01), SVM(alpha=1), MultinomialNB(alpha=0.001)       |   1.16   |               38.64
|
|      TFIDF      |                Naive Bayes                 | SelectKBest top 1000 |
alpha=1                                     |   1.17   |               40.6                  |
|      TFIDF      | Logistic Regression + Class Balancing      | SelectKBest top 1000 |
alpha=0.0001                                |   1.09   |               39.09                 |
|      TFIDF      |                Linear SVM                  | SelectKBest top 1000 |
alpha=0.0001                                |   1.16   |               38.53                 |
|      TFIDF      |             Random Forest                  | SelectKBest top 1000 |
n_estimator=2000, max_depth=10              |   1.26   |               48.68                 |
```

```
+----------------+----------------------------------------+---------------------+-----------------
-------------------------------------------------+----------+-------------------------------------+
|      TFIDF     |            Stacking Models             | SelectKBest top 1000 | LR[LR(alpha=0.01
, SVM(alpha=1), MultinomialNB(alpha=0.001)]alpha=0.1 |   1.22   |               41.35
|
|      TFIDF     |             Maximum Voting             | SelectKBest top 1000 |
LR(alpha=0.01), SVM(alpha=1), MultinomialNB(alpha=0.001)       |   1.2    |               38.79
|
|   Bag of Words  | Logistic Regression + Class Balancing |   Unigram +Bigram    |
alpha=0.01                              |   1.21   |               38.72               |
|   Bag of Words  | Logistic Regression + Class Balancing |   (1,2,3,4) -grams   |
alpha=0.0001                             |  1.005   |               34.58               |
+----------------+----------------------------------------+---------------------+-----------------
-------------------------------------------------+----------+-------------------------------------+
```

Steps that were followed during this case study

1. Initially the type of data was analysed, with all the constraints that it brings to the table and the information availability that is there.
2. Initially to understand the data better Exploratory data analysis was performed on each feature
3. In the EDA important questions were asked for all the three features :---

```
- What type of feature is it?
- How good is that feature?
- How do we actually featurize that feature?
- Is that feature stable?
```

4. Machine Learning models are applied on top of that firsly for simple Bag of words featurization on the available features.
5. Then these models were again applied, but this time on top of TFIDF vectorization and the results were taken into consideration.
6. After that top 1000 features from the tfidf vectorization were taken and the models were applied on them.
7. Going much further we tae unigrams and bigrams of the text feature and apply Logistic regression on top of that the vectors after stacking them.
8. Then to decrease the log-loss more I actually take up 1-6 grams and then from that took only top 3000 features using the selectKBest feature selection technique and applied Logistic regression on top of that actually helped us reduce the logarithmic loss up to just 1.003.