CSE 158

HW2

1. The code currently does not perform any train/test splits. Split the data into training, validation, and test sets, via 1/3, 1/3, 1/3 splits. Use the first third, second third, and last third of the data (respectively). After training on the training set, report the accuracy of the classifier on the validation and test sets (1mark).

valid acc: 0.90028199436

test acc: 0.577788444231

2. Let's come up with a more accurate classifier based on a few common words in the review. Build a feature vector to implement a classifier of the form p(positive label) $=\sigma(\theta_0+\theta_1\times$#'lactic' $+\theta_2\times$#'tart'...),where each feature corresponds to the number of times a particular word appears. Base your feature on the following 10 words: "lactic," "tart," "sour," "citric," "sweet," "acid," "hop," "fruit," "salt," "spicy." Convert the reviews to lowercase before counting.

Function code is in below section

3. Report the number of true positives, true negatives, false positives, false negatives, and the Balanced Error Rate of the classifier on the test set (1 mark).

Used Q2 features

```
Question 3:
TN: 290.0
TP: 5806.0
FN: 106.0
FP: 10465.0
balanced error rate:  0.495482715972
```

4. Implement a training/validation/test pipeline so that you can select the best model based on its performance on the validation set. Try models with $\lambda \in \{0,0.01,0.1,1,100\}$. Report the performance on the training/validation/test sets for the best value of λ (1 mark).

Used features in Question 2

```
best lamda:  100
acc on train: 0.561442457698
acc on valid: 0.95122097558
acc on test: 0.357472850543
```

5. Find and report the PCA components (i.e., the transform matrix) using the week 3 code (1 mark).

```
PCA components on train set:
[[  2.91553267e-04   3.36316078e-03  -4.92016461e-03   1.22605289e-02
    8.02994923e-01   2.01250020e-04   5.90516350e-01   7.22369875e-02
    1.75711994e-04   3.29456581e-02]
 [ -2.36616086e-03  -8.55889699e-03  -1.91717318e-02   1.45794087e-02
   -5.92410182e-01   4.74330032e-04   8.05130166e-01  -4.97366274e-03
   -1.28579556e-03   1.14021305e-02]
 [  3.99780701e-03   4.58510040e-02   1.01611015e-01   1.87044464e-03
   -6.24232301e-02  -5.23321160e-05  -3.73196976e-02   9.90698551e-01
    9.18564879e-04   2.79198868e-02]
 [ -1.52012269e-04   2.00218881e-02  -1.04163870e-02   2.44992682e-02
   -1.80912429e-02   8.60712007e-05  -2.81132009e-02  -3.02427712e-02
    2.45592725e-03   9.98424798e-01]
 [  2.55382520e-02   2.24799457e-01   9.67585118e-01   7.19960887e-03
   -2.31187652e-03   9.62965425e-03   2.29568098e-02  -1.09118707e-01
    9.38463609e-04   2.70997623e-03]
 [  3.54769091e-02   9.72027558e-01  -2.29341823e-01   1.34748385e-02
   -4.20493294e-03   8.77575267e-03   1.83813401e-03  -2.11898082e-02
    7.29985353e-03  -2.28954543e-02]
 [  5.66835271e-03  -1.55377546e-02  -3.62210660e-03   9.99363327e-01
   -5.81954419e-04   5.54584002e-03  -1.84164136e-02  -8.50413772e-04
    2.59682946e-04  -2.48036199e-02]
 [  9.96955075e-01  -4.09453361e-02  -1.74561630e-02  -6.65912506e-03
```

```
  -1.16153723e-03   6.34885323e-02   1.28440550e-03  -3.69526311e-04
  -4.48757759e-03   9.63660068e-04]
 [ -6.40484881e-02  -7.96385717e-03  -6.17769008e-03  -5.32627798e-03
   2.60276502e-04   9.97852694e-01  -7.26121605e-04   1.31450704e-03
  -7.44873354e-03   1.72584485e-04]
 [  3.70628217e-03  -7.64907081e-03   5.51204371e-04  -4.79523814e-04
  -7.71354670e-04   7.64404986e-03   1.00504146e-03  -5.89435260e-04
   9.99930789e-01  -2.29235033e-03]]
```

6. How would the first data point (i.e., X[0]) be represented in the PCA basis you have found (1 mark)?

```
[  3.94626731e-03  -1.32912187e-02   5.38466180e-02   1.97178635e-02
  2.75875961e-01   1.04298138e+00  -4.20104922e-03   1.95296481e+00
 -1.36060833e-01  -2.36506462e-04]
```

7. Suppose we want to compress the data using just two PCA dimensions. How large is the reconstruction error when doing so (1 mark)?

```
Question 7:
Reconstruction error of d = 2:
covariance Matrix eig val:  8346.76973501
```

8. Looking at the first two dimensions of our data in the PCA basis is an effective way to ‘summarize’ the data via a 2-d plot. Using a plotting program of your choice, make a 2-d scatterplot showing the difference between ‘American IPA’ style beers versus all other styles (e.g. plot American IPAs in red and other styles in blue) (1 mark).