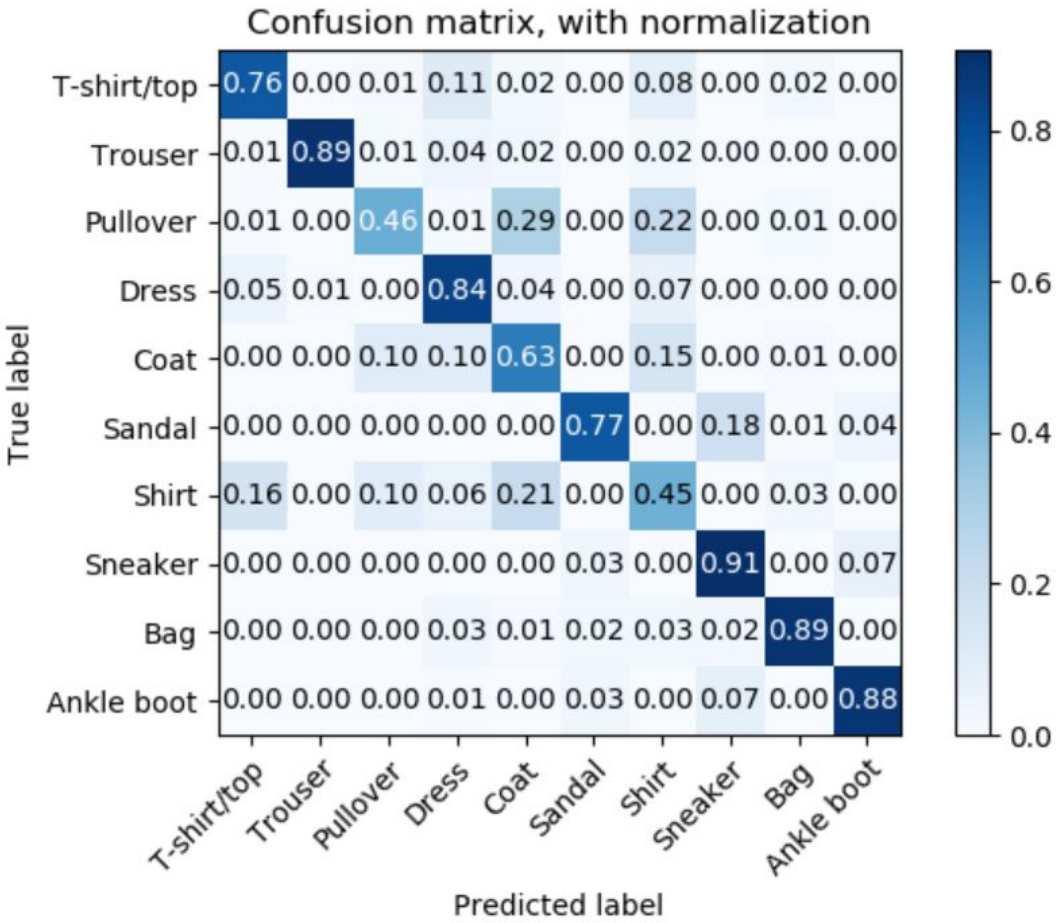
Henry Solberg (R4) and Ayush Sinha (Q4)
CS 440 MP3
3/31/2019

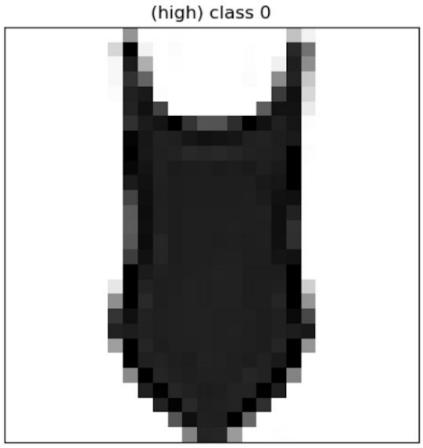
**Section I:**

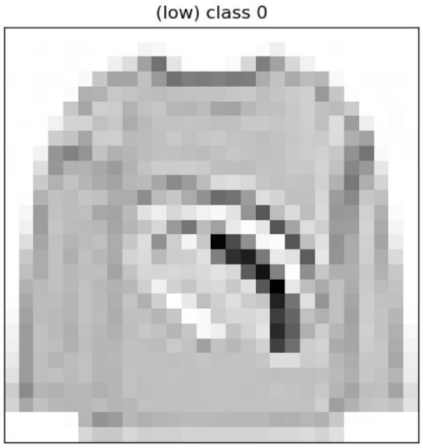
We implemented the assigned methods for a naive bayes and a perceptron machine learning algorithm.

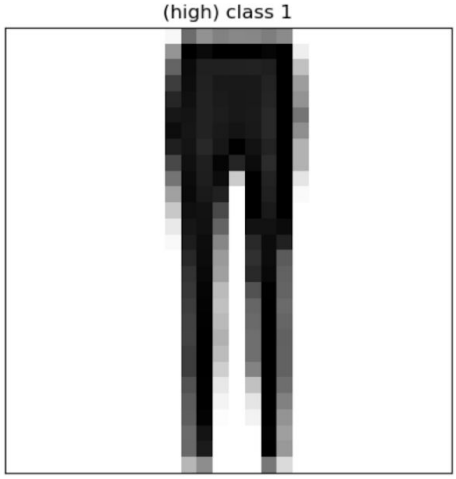
- Naive Bayes
  - Average classification rate: 74.7%
  - Average classification rate per class:
    - T-shirt/top: 76%
    - Trouser: 89%
    - Pullover: 46%
    - Dress: 84%
    - Coat: 63%
    - Sandal: 77%
    - Shirt: 45%
    - Sneaker: 91%
    - Bag: 89%
    - Angle boot: 88%
  - Confusion matrix:

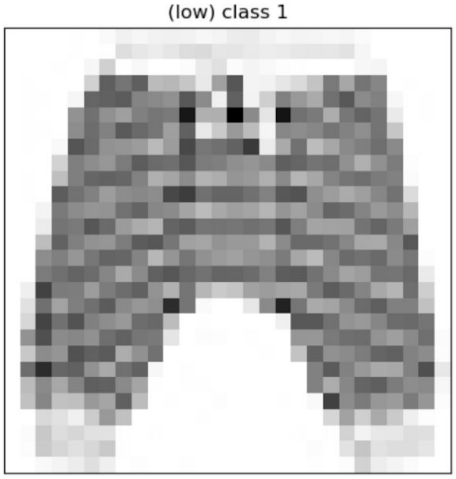


- Cont…
    - We see that pullovers, coats, and shirts have the lowest classification rate, and furthermore are most often confused with each other. In other words, the confusion matrix of just these three classes would be rather poor: each cell would have relatively large values. Examining the visualization below, we see that, indeed, pullovers, coats, and shirts have similar shapes. Intuitively, from a human perspective pullovers, coats, and shirts are differentiated more by features like buttons, fabric/material, and thickness and so it seems reasonable that our shape/brightness-based naive bayes classifier would not distinguish these three classes very well.
    - Test examples of high and low posterior probability:

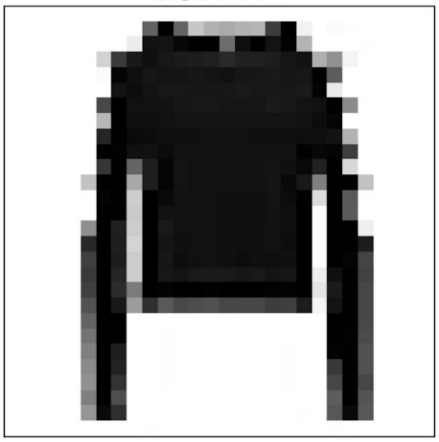


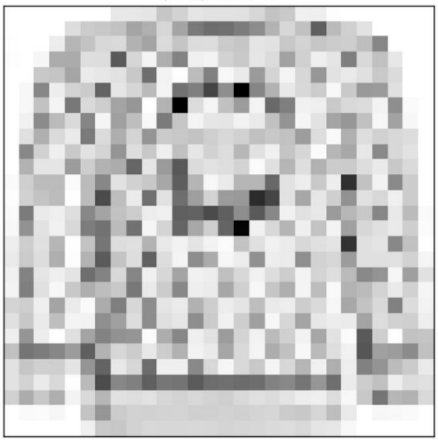



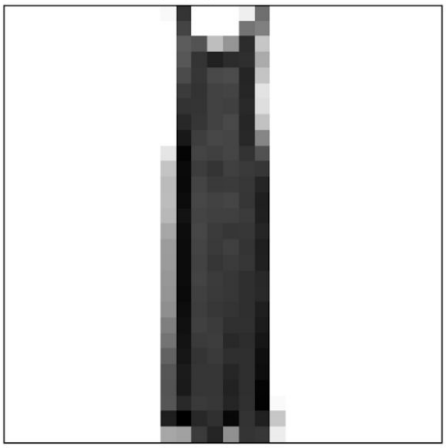
(high) class 2



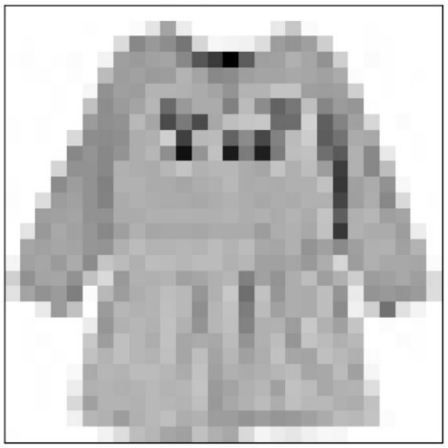
(low) class 2



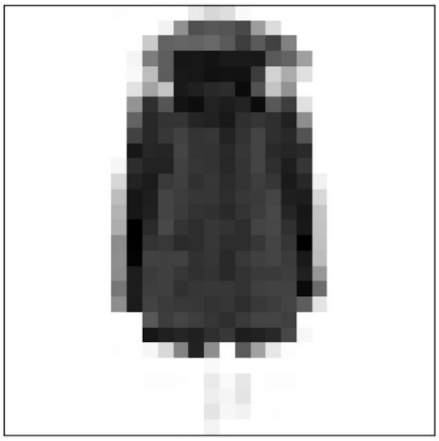
(high) class 3



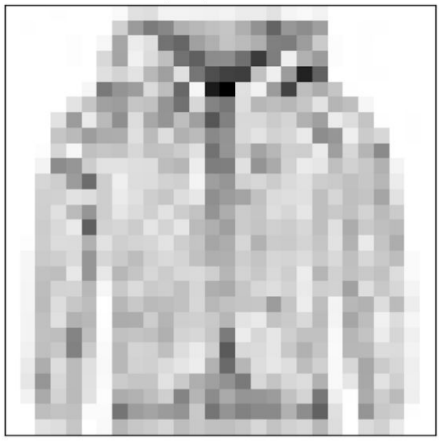
(low) class 3



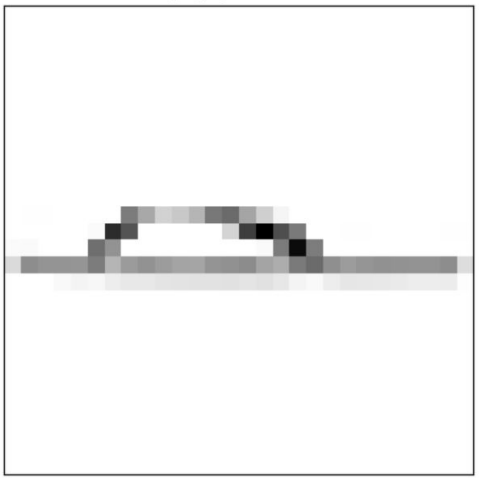
(high) class 4



(low) class 4

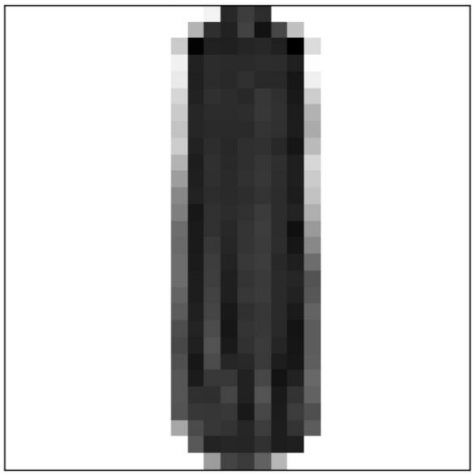


(high) class 5

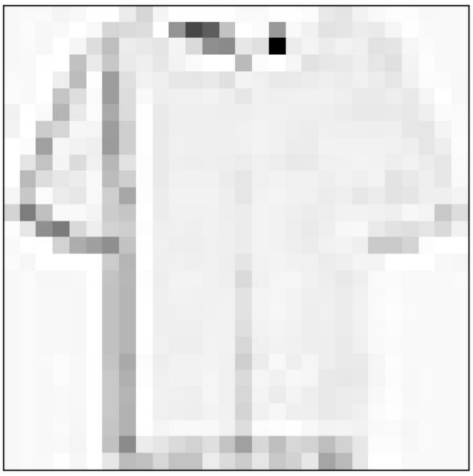


(low) class 5

(high) class 6



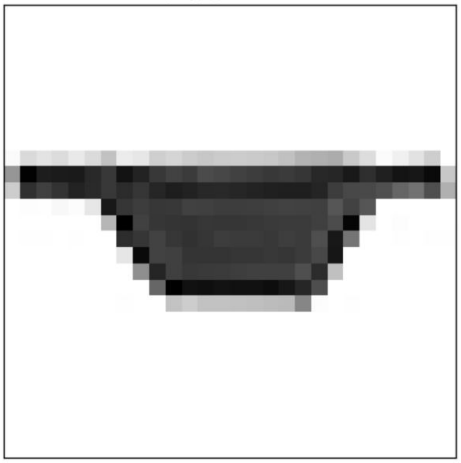
(low) class 6



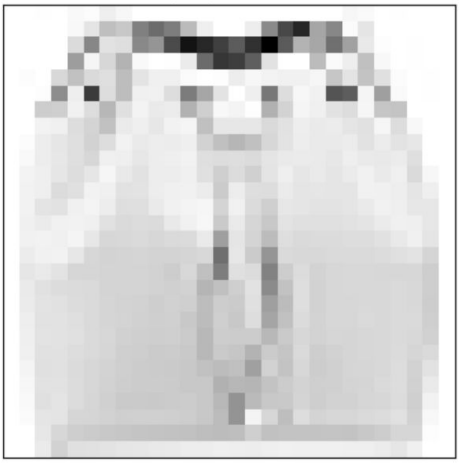
(high) class 7

(low) class 7

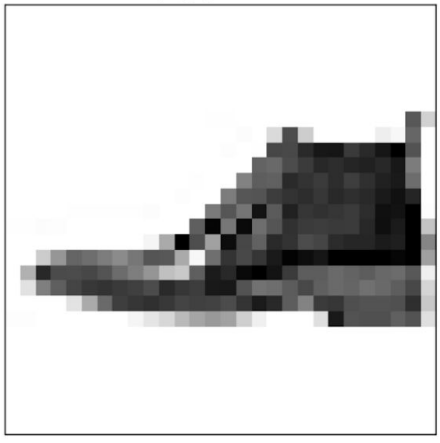
(high) class 8



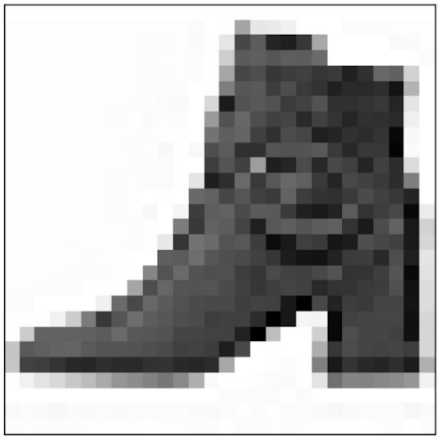
(low) class 8



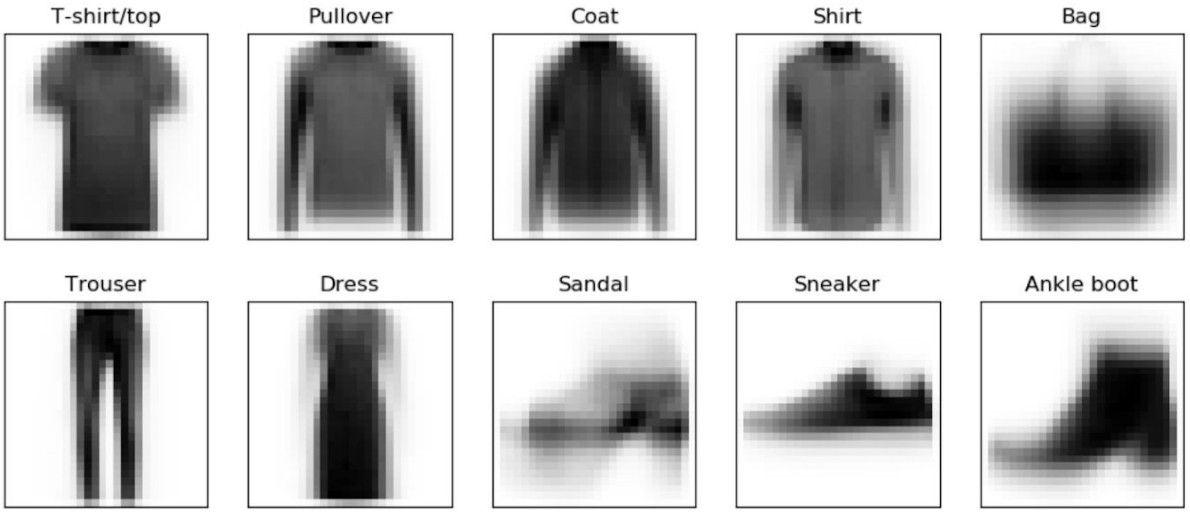
(high) class 9



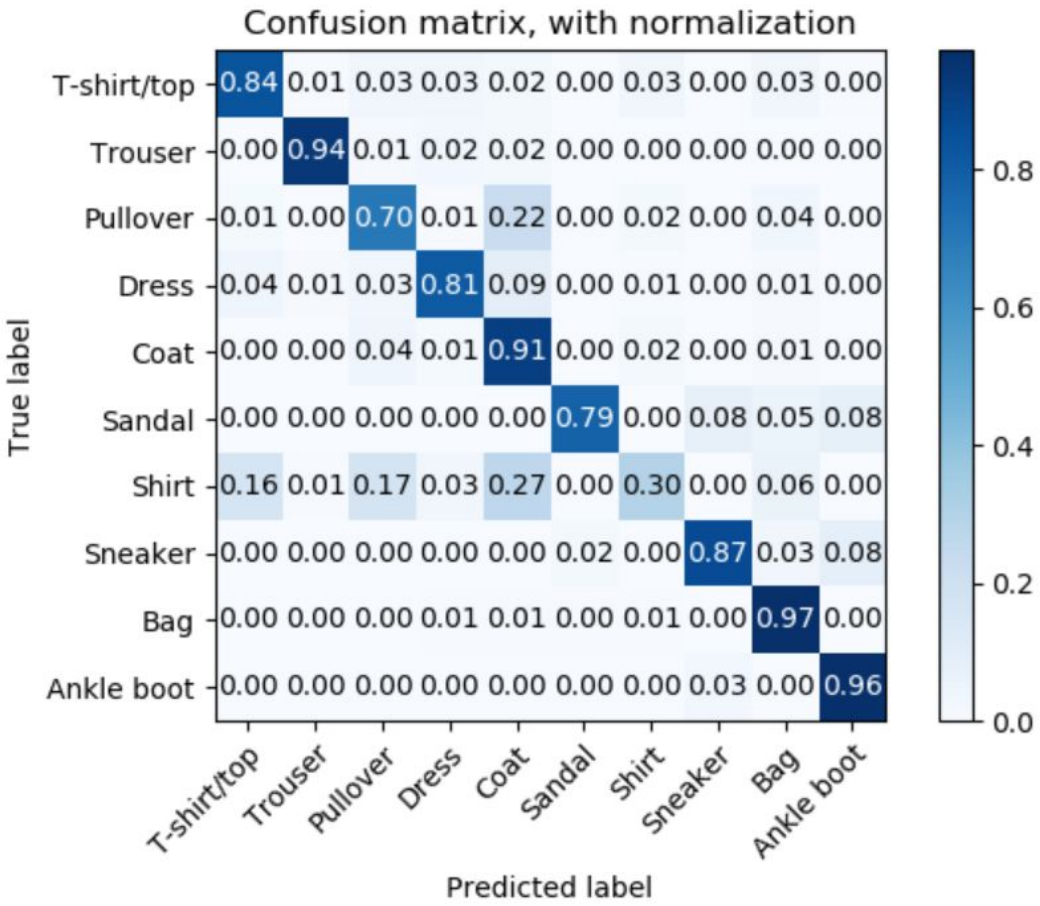
(low) class 9



- Cont…
    - We see that the items with greatest posterior probability are often dark items that occupy only the most central areas of the torso (or foot or leg) part of the item, whereas items with low posterior probability are often light and/or spread out. We speculate that this is because darker items are more common in the training set and positions of peripheral parts of the items are unreliable, whereas most items have a central piece shared between most of the images of that class.
    - Weight visualization:

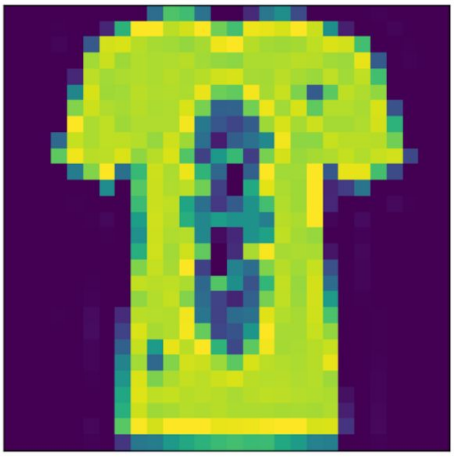



- Perceptron
    - Average classification rate: 79.83%
    - Average classification rate per class:
        T-shirt/top: 84%
        Trouser: 94%
        Pullover: 70%
        Dress: 81%
        Coat: 91%
        Sandal: 79%
        Shirt: 30%
        Sneaker: 87%
        Bag: 97%
        Angle boot: 96%
    - Confusion matrix:

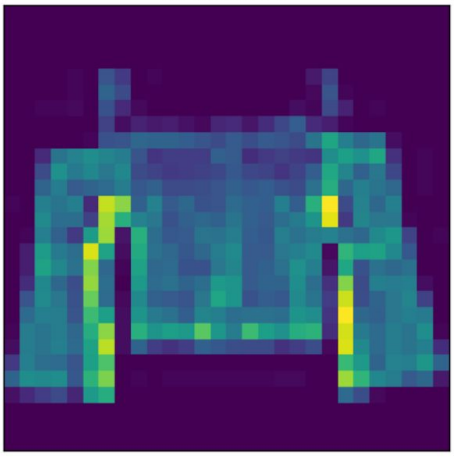



- Cont…
  - We see shirts are often classified as coats, pullovers, or t-shirts. Again we speculatively attribute this to the classifier making decisions based on color and shape as opposed to material, buttons, etc.. Intuitively speaking, it would be very hard to distinguish, for example, a button-down shirt from a pullover based on shape and color alone.
  - Test examples of high and low perceptron score:

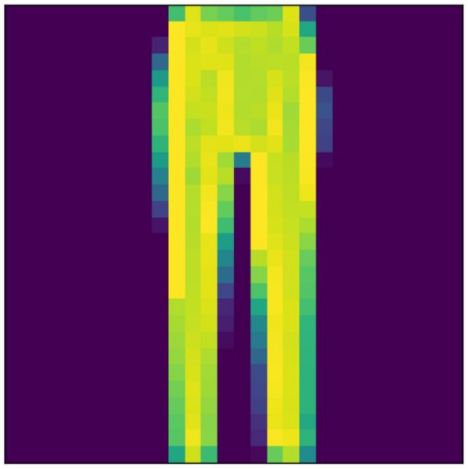
(high) class 0



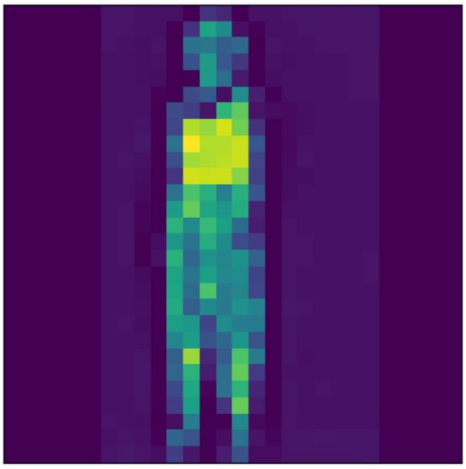
(low) class 0



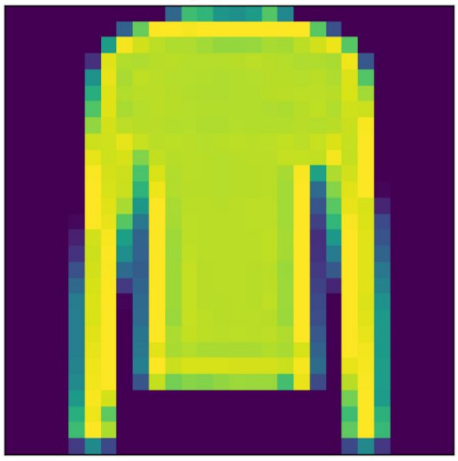
(high) class 1



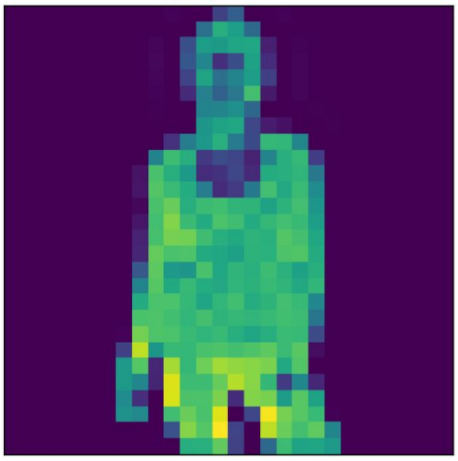
(low) class 1



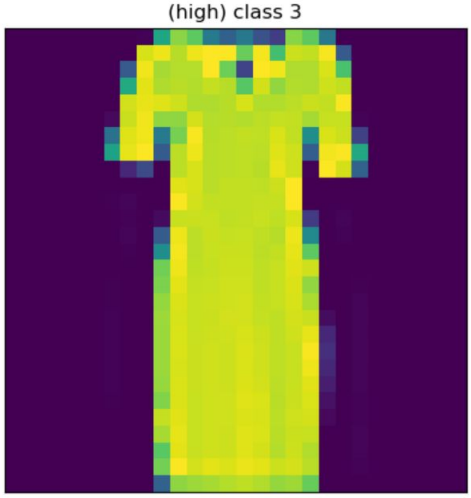
(high) class 2



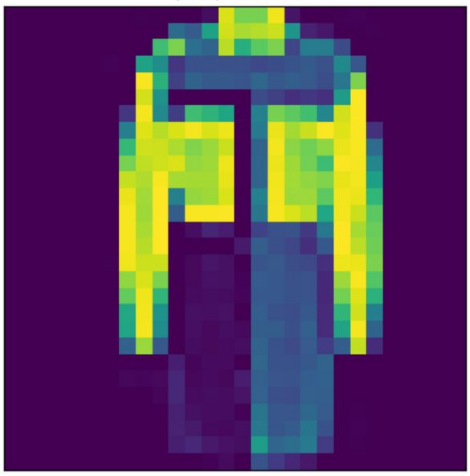
(low) class 2



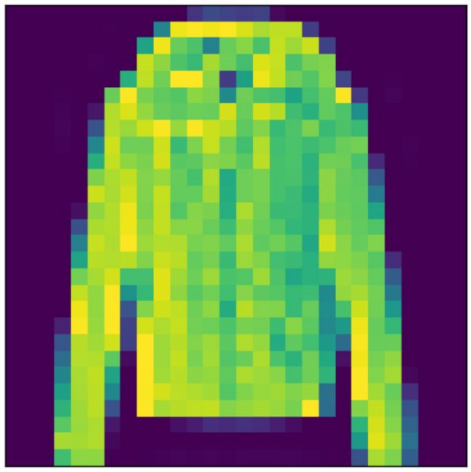
(high) class 3



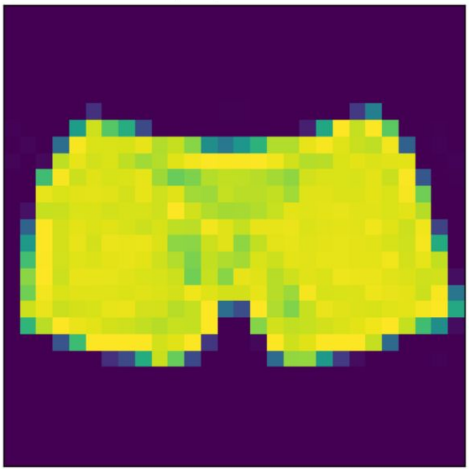
(low) class 3



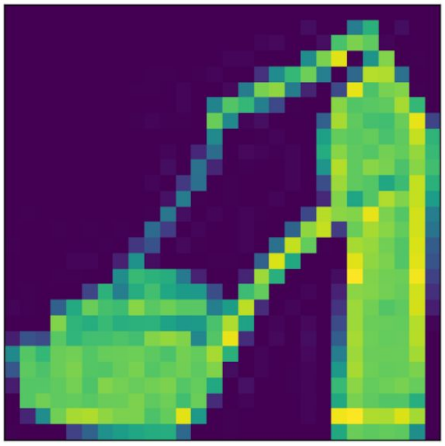
(high) class 4



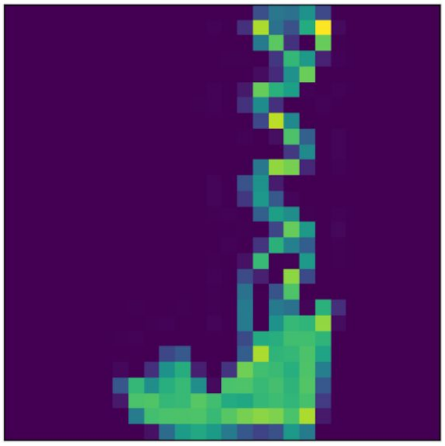
(low) class 4



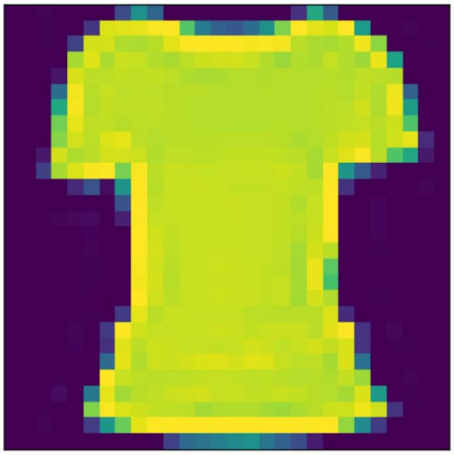
(high) class 5



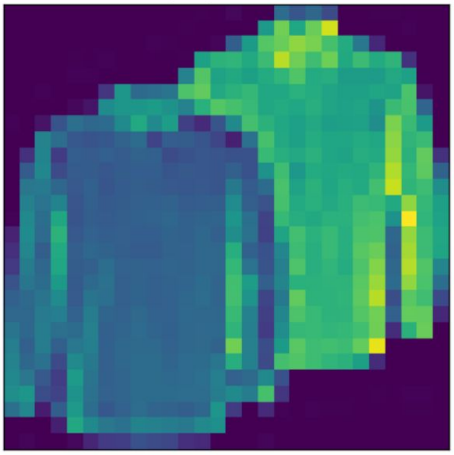
(low) class 5



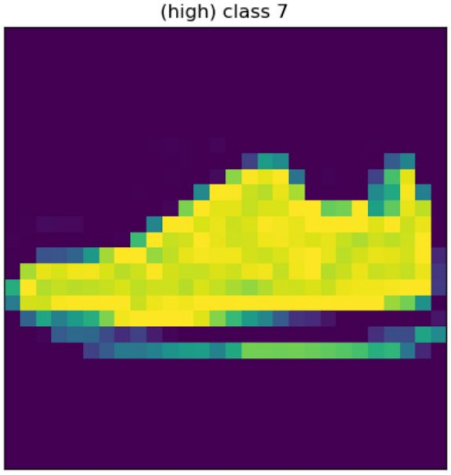
(high) class 6



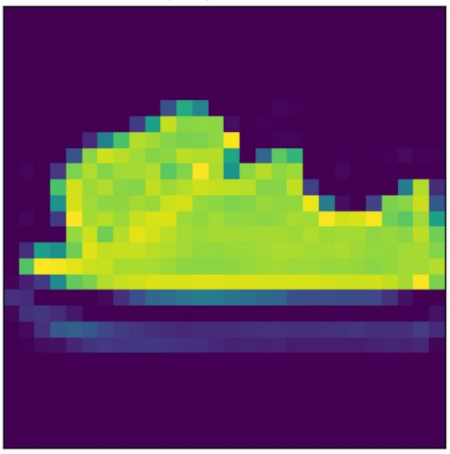
(low) class 6



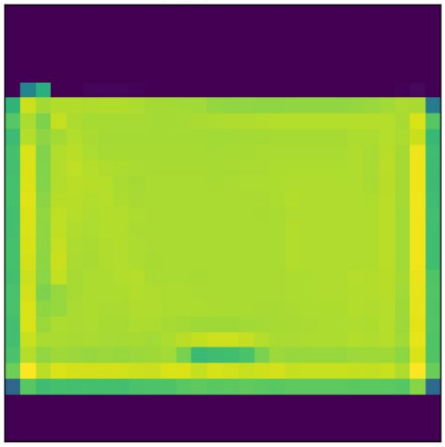
(high) class 7



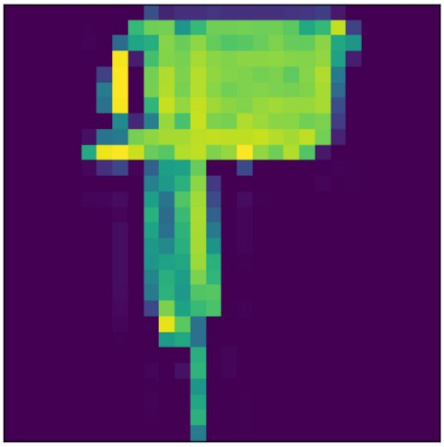
(low) class 7



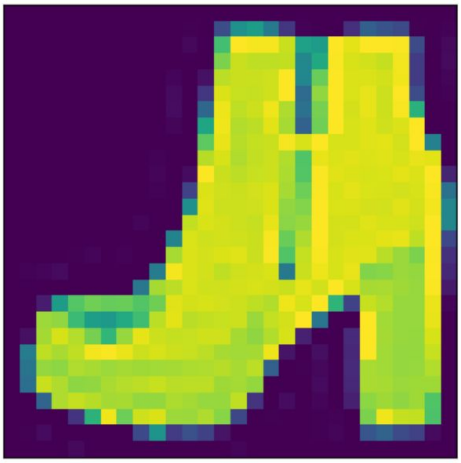
(high) class 8



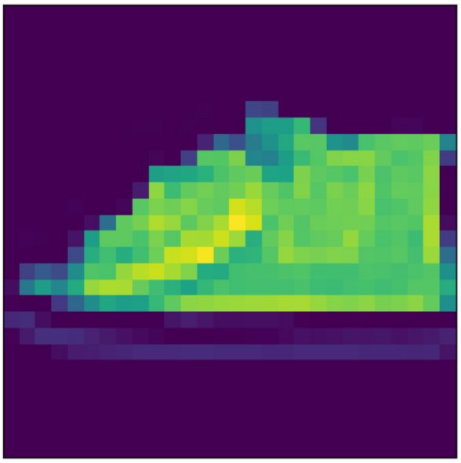
(low) class 8



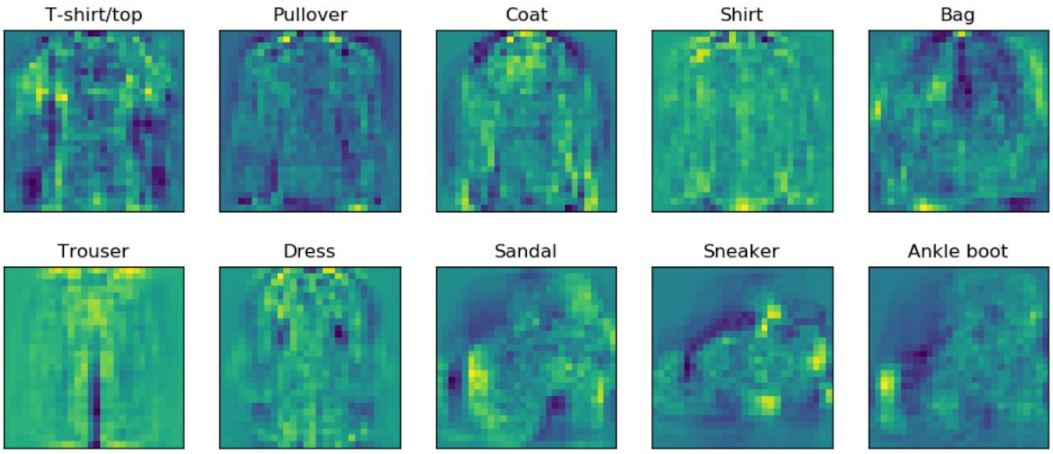
(high) class 9



(low) class 9



- Cont...
    - We see examples with a low perceptron score are highly atypical of their class. Some include human models, or even two items in one picture.
    - Motivated by the below visualization, we decided to use only one pass through the training data (one epoch). If we do, for example, two epochs, the below visualization starts to make a lot less sense. Many peripheral pixels became highly relevant for specific classes, which we speculate is a case of overfitting (for example, some corner pixel happened to have a particular value in several T-shirt images in the training set, but that is simply a coincidence and that pixel does not make an image more likely to be a t-shirt).
    - Perceptron visualization:



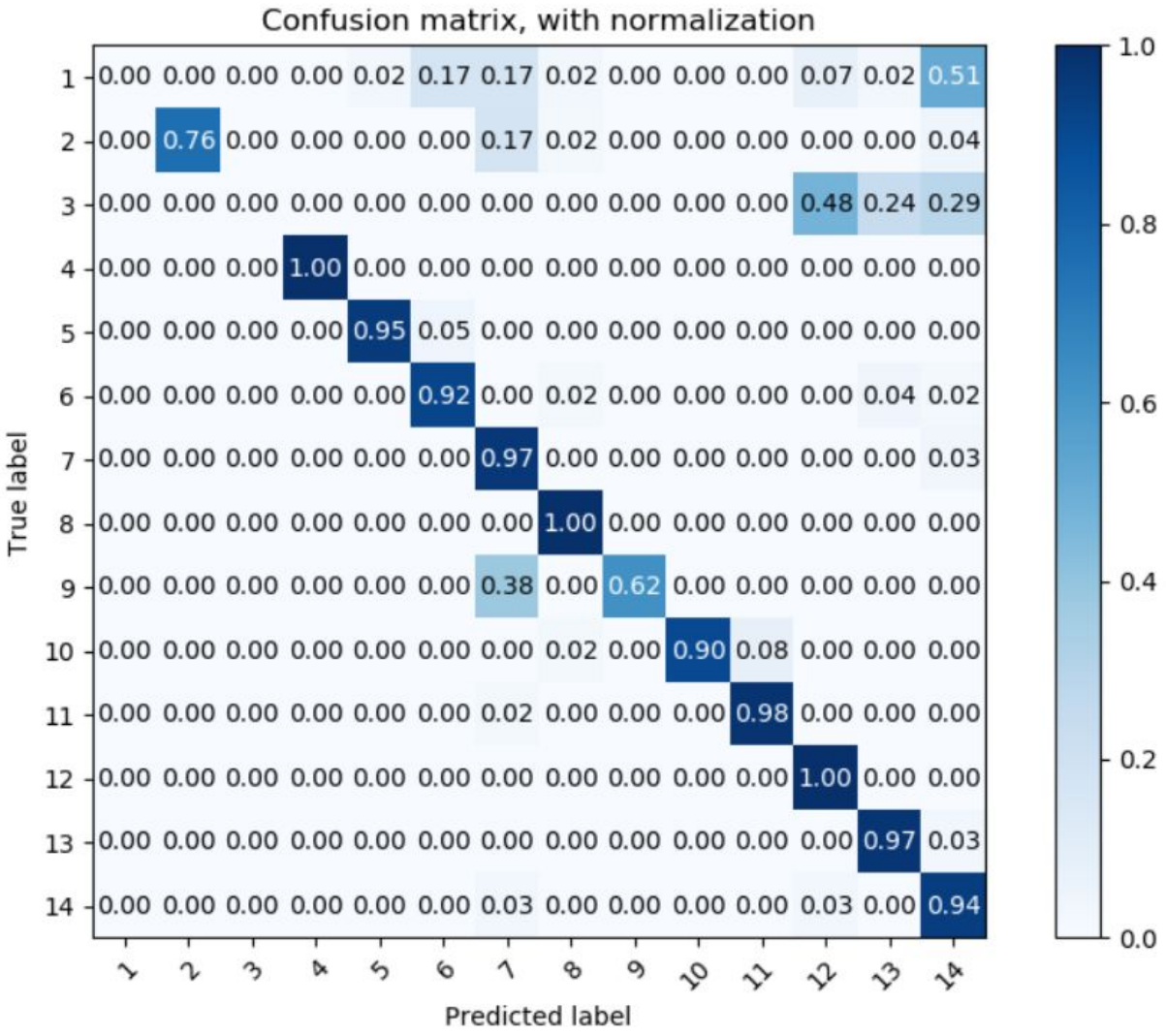

## Section II Text Classification

We constructed a Bag of Words model and learnt a Naive Bayes classifier.
In fit() function, we go through all words in train_set and compute their likelihoods.
The likelihood of a word given class is defined as number of occurrences of that word in documents of given class per number of total number of words in documents of given class.
Laplace smoothing is also done to handle likelihoods for words that occur in train_set but might not occur in some classes.
The vocabulary is stored in a set. Vocabulary refers to the set of all words seen in the train_set.
Prior probabilities for each class is also calculated as number of documents of a class per total number of documents.

In predict(), we go through each list of words in x_set, one at a time, and calculate total likelihood of all words for all classes. The total likelihood is simply the sum of individual likelihoods (log taken during training). Next, we add prior probability of the class to get posterior probabilities. The class corresponding to maximum posterior probability is set as predicted class and saved in result list.

Overall accuracy is computed as number of correctly labelled lists in x_set per total number of lists in x_set.

Confusion matrix, recall, precision and F1 scores are presented below (for test with priors)




**Precision for all classes :** [0.0, 1.0, 0.0, 1.0, 0.9545454545454546, 0.8461538461538461, 0.5918367346938775, 0.8947368421052632, 1.0, 1.0, 0.9166666666666666, 0.75, 0.8222222222222222, 0.5076923076923077]

**Recall for all classes:** [0.0, 0.7608695652173914, 0.0, 1.0, 0.9545454545454546, 0.9166666666666666, 0.9666666666666667, 1.0, 0.625, 0.9, 0.9777777777777777, 1.0, 0.9736842105263158, 0.9428571428571428]

**F1 Score for all classes:** [0.0, 0.8641975308641976, 0.0, 1.0, 0.9545454545454546, 0.8799999999999999, 0.7341772151898733, 0.9444444444444444, 0.7692307692307693, 0.9473684210526316, 0.946236559139785, 0.8571428571428571, 0.891566265060241, 0.6599999999999999]

**Accuracy** 0.8116

**The top 20 most likely words for each class are:**
1
['company', 'based', 'business', 'founded', 'records', 'bergen', 'record', 'services', 'systems', 'office', 'products', 'distribution', 'college', 'buses', 'also', 'regional', 'national', 'university', 'sports', 'including']
10
['family', 'species', 'found', 'genus', 'moth', 'gastropod', 'sea', 'known', 'marine', 'described', 'tropical', 'snail', 'mollusk', 'endemic', 'subtropical', 'habitat', 'natural', 'forests', 'snails', 'moist']
14
['published', 'book', 'novel', 'first', 'journal', 'written', 'series', 'newspaper', 'story', 'american', 'author', 'new', 'magazine', 'fiction', 'books', 'peerreviewed', 'also', 'science', 'publication', 'life']
13
['film', 'directed', 'starring', 'american', 'stars', 'released', 'written', 'based', 'drama', 'comedy', 'also', 'produced', 'films', 'silent', 'first', 'movie', 'roles', 'novel', 'name', 'documentary']
12
['album', 'released', 'band', 'records', 'first', 'studio', 'american', 'songs', 'music', 'second', 'release', 'recorded', 'rock', 'debut', 'live', 'tracks', 'label', 'albums', 'new', 'ep']
6
['navy', 'built', 'war', 'ship', 'uss', 'united', 'class', 'aircraft', 'world', 'states', 'launched', 'service', 'designed', 'named', 'first', 'royal', 'commissioned', 'american', 'ii', 'company']
4
['born', 'football', 'played', 'league', 'plays', 'professional', 'player', 'footballer', 'former', 'national', 'american', 'also', 'currently', 'hockey', 'rugby', 'team', 'australian', 'november', 'world', 'new']
2
['school', 'high', 'located', 'university', 'college', 'schools', 'public', 'students', 'education', 'district', 'county', 'new', 'founded', 'one', 'independent', 'city', 'part', 'established', 'united', 'private']
7
['historic', 'house', 'built', 'located', 'church', 'building', 'national', 'register', 'places', 'listed', 'county', 'street', 'united', 'known', 'also', 'museum', 'states', 'designed', 'hospital', 'added']
8
['river', 'lake', 'mountain', 'located', 'south', 'km', 'north', 'county', 'near', 'tributary', 'west', 'range', 'lies', 'creek', 'crater', 'east', 'state', 'ft', 'flows', 'pass']
11
['species', 'family', 'plant', 'genus', 'native', 'endemic', 'flowering', 'known', 'found', 'common', 'leaves', 'plants', 'habitat', 'tree', 'name', 'grows', 'orchid', 'south', 'bulbophyllum', 'perennial']

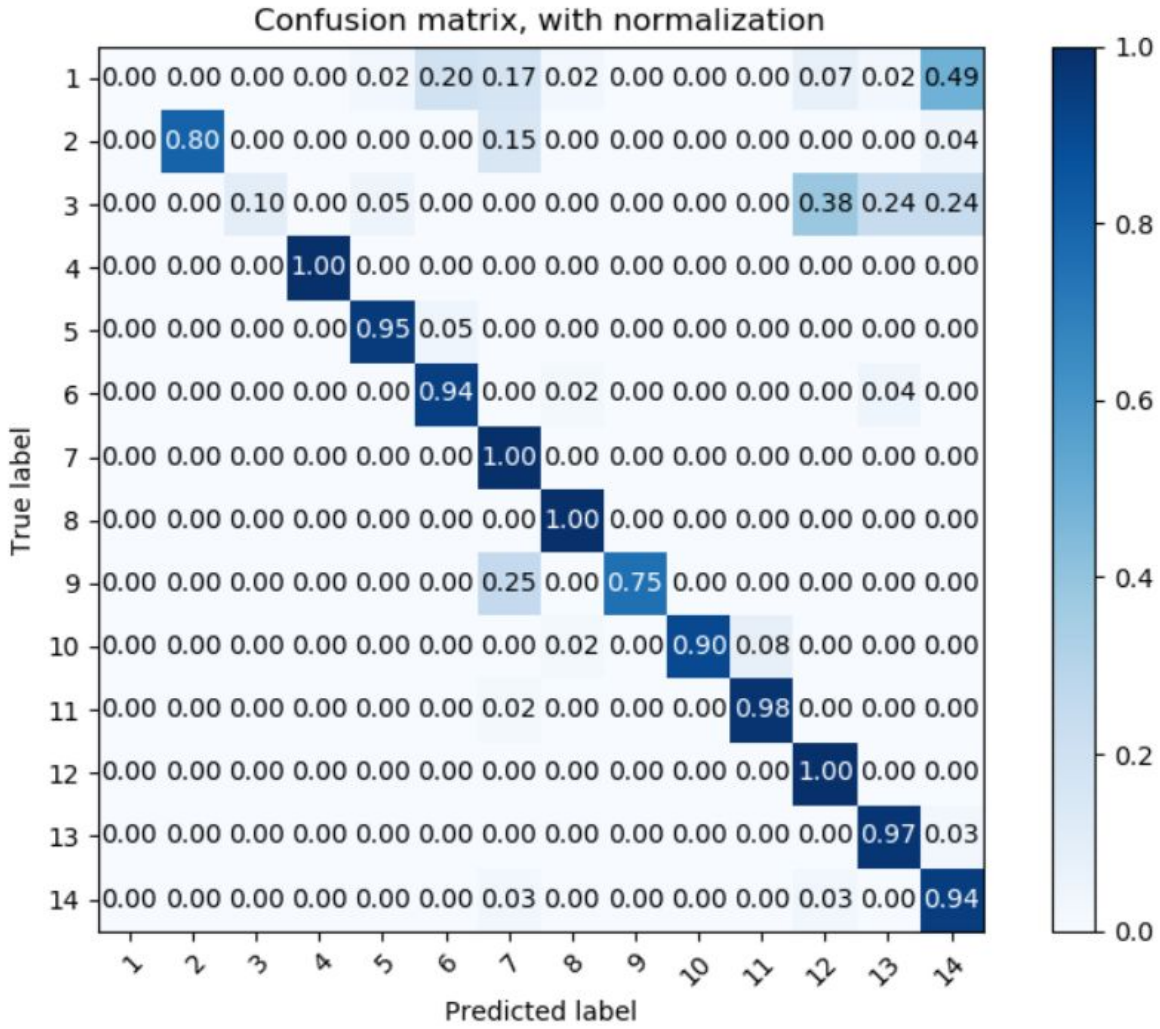
3
['born', 'american', 'known', 'new', 'band', 'best', 'writer', 'rock', 'work', 'musician', 'music', 'also', 'singer', 'york', 'author', 'album', 'books', 'university', 'member', 'united']
5
['born', 'member', 'district', 'politician', 'state', 'democratic', 'house', 'senate', 'party', 'served', 'former', 'county', 'since', 'representatives', 'republican', 'elected', 'united', 'american', 'national', 'representing']
9
['village', 'district', 'population', 'province', 'located', 'census', 'municipality', 'nepal', 'india', 'state', 'county', 'people', 'km', 'within', '2010', '1991', 'time', 'kerala', 'south', 'zone']

The scores for class 1 and 3 are very low. This is because no words in classes 1 or 3 had very high likelihood in training data, but words that do occur in class 1 and 3 in test data have high likelihood in other classes. For example, documents belonging to class 3 are often misclassified as belonging to class 12. Investigating further, we see the most likely words for class three (born, american, known) all have log(probability) below -6.0. We see the most common words in test data for class three are born, known, american, and music. The words american and music, however, have log(probability) around -5.9 in training data for class *twelve*.

A similar phenomenon is observable in class 1, where the two most likely words have log(probability) -6.7 and -7.7 (very low!). Even “meaningless” words such as “also” have higher probability in other classes (e.g. class 14), leading to misclassification of documents from class 1. We therefore conclude that our classifier’s poor performance on documents from class 1 and 3 is an expected result of our choice of model.

**Next, we test the effect of using ML (maximum likelihood) instead of MAP.** We simply omitted the addition of prior probability to total likelihood and ran our code again.

Confusion matrix, recall, precision and F1 scores are presented below (ML)




**Precision for all classes :** [0.0, 1.0, 1.0, 1.0, 0.9130434782608695, 0.8333333333333334, 0.625, 0.918918918918919, 1.0, 1.0, 0.9166666666666666, 0.7777777777777778, 0.8222222222222222, 0.5409836065573771]

**Recall for all classes:** [0.0, 0.8043478260869565, 0.09523809523809523, 1.0, 0.9545454545454546, 0.9375, 1.0, 1.0, 0.75, 0.9, 0.9777777777777777, 1.0, 0.9736842105263158, 0.9428571428571428]

**F1 Score for all classes:** [0.0, 0.891566265060241, 0.17391304347826084, 1.0, 0.9333333333333332, 0.8823529411764706, 0.7692307692307693, 0.9577464788732395, 0.8571428571428571, 0.9473684210526316, 0.946236559139785, 0.8750000000000001, 0.891566265060241, 0.6875]

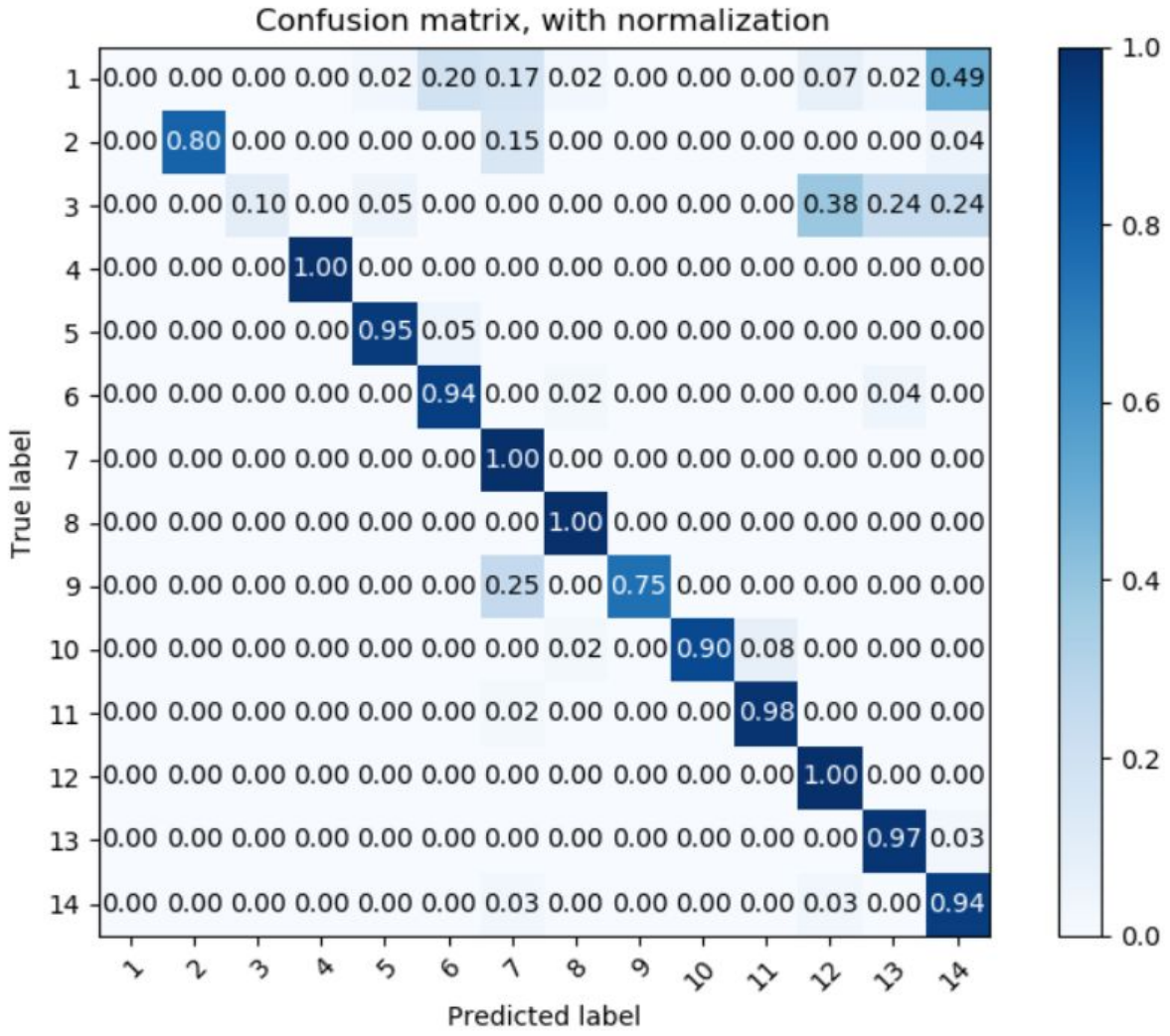
**Accuracy** 0.8261

**The top 20 most likely words for each class are:**
1
['company', 'based', 'business', 'founded', 'records', 'record', 'bergen', 'services', 'systems', 'products', 'office', 'school', 'capel', 'inc', 'norwegian', 'sports', 'health', 'university', 'regional', 'toronto']
10
['family', 'species', 'found', 'genus', 'moth', 'gastropod', 'sea', 'known', 'marine', 'described', 'tropical', 'snail', 'mollusk', 'endemic', 'subtropical', 'habitat', 'natural', 'forests', 'snails', 'moist']
14
['published', 'book', 'novel', 'first', 'journal', 'written', 'series', 'newspaper', 'story', 'american', 'author', 'new', 'magazine', 'fiction', 'books', 'peerreviewed', 'also', 'science', 'publication', 'life']
13
['film', 'directed', 'starring', 'american', 'stars', 'released', 'written', 'based', 'drama', 'comedy', 'produced', 'also', 'films', 'silent', 'first', 'movie', 'roles', 'name', 'novel', 'documentary']
12
['album', 'released', 'band', 'records', 'first', 'studio', 'american', 'songs', 'music', 'second', 'release', 'recorded', 'rock', 'debut', 'live', 'tracks', 'label', 'albums', 'new', 'ep']
6
['navy', 'built', 'war', 'ship', 'uss', 'united', 'aircraft', 'class', 'world', 'states', 'launched', 'service', 'named', 'designed', 'first', 'royal', 'commissioned', 'american', 'ii', 'us']
4
['born', 'football', 'played', 'league', 'player', 'professional', 'plays', 'footballer', 'former', 'national', 'american', 'currently', 'also', 'hockey', 'rugby', 'team', 'australian', 'november', 'world', 'new']
2
['school', 'high', 'located', 'university', 'college', 'public', 'schools', 'students', 'education', 'county', 'district', 'founded', 'new', 'one', 'part', 'city', 'united', 'independent', 'established', 'catholic']
7
['historic', 'house', 'built', 'located', 'church', 'building', 'national', 'register', 'places', 'listed', 'county', 'street', 'united', 'known', 'also', 'museum', 'states', 'designed', 'added', 'hospital']

8
['river', 'lake', 'mountain', 'located', 'south', 'km', 'north', 'county', 'near', 'tributary', 'range', 'west', 'lies', 'creek', 'east', 'crater', 'state', 'ft', 'flows', 'pass']
11
['species', 'family', 'plant', 'genus', 'native', 'endemic', 'flowering', 'known', 'found', 'common', 'leaves', 'plants', 'habitat', 'tree', 'name', 'grows', 'orchid', 'south', 'bulbophyllum', 'perennial']
3
['born', 'american', 'known', 'new', 'band', 'writer', 'best', 'rock', 'music', 'work', 'musician', 'singer', 'also', 'york', 'author', 'books', 'album', 'united', 'university', 'guitarist']
5
['born', 'member', 'district', 'politician', 'state', 'house', 'democratic', 'senate', 'party', 'served', 'former', 'county', 'since', 'representatives', 'republican', 'united', 'elected', 'american', 'representing', 'national']
9
['village', 'district', 'population', 'province', 'located', 'census', 'municipality', 'nepal', 'india', 'state', 'county', 'km', 'people', 'within', '1991', '2010', 'township', 'central', 'southern', 'zone']

**Last, we tested the effect of using a uniform distribution for our class prior calculation**. The class priors were all set to 1/14.

Confusion matrix, recall, precision and F1 scores are presented below (uniform priors)




**Precision for all classes :** [0.0, 1.0, 1.0, 1.0, 0.9130434782608695, 0.8333333333333334, 0.625, 0.918918918918919, 1.0, 1.0, 0.9166666666666666, 0.7777777777777778, 0.8222222222222222, 0.5409836065573771]

**Recall for all classes:** [0.0, 0.8043478260869565, 0.09523809523809523, 1.0, 0.9545454545454546, 0.9375, 1.0, 1.0, 0.75, 0.9, 0.9777777777777777, 1.0, 0.9736842105263158, 0.9428571428571428]

**F1 Score for all classes:** [0.0, 0.891566265060241, 0.17391304347826084, 1.0, 0.9333333333333332, 0.8823529411764706, 0.7692307692307693, 0.9577464788732395, 0.8571428571428571, 0.9473684210526316, 0.946236559139785, 0.8750000000000001, 0.891566265060241, 0.6875]

**Accuracy** 0.8261

**The top 20 most likely words for each class are:**

1
['company', 'based', 'business', 'founded', 'records', 'record', 'bergen', 'services', 'systems', 'products', 'office', 'school', 'capel', 'inc', 'norwegian', 'sports', 'health', 'university', 'regional', 'toronto']

10
['family', 'species', 'found', 'genus', 'moth', 'gastropod', 'sea', 'known', 'marine', 'described', 'tropical', 'snail', 'mollusk', 'endemic', 'subtropical', 'habitat', 'natural', 'forests', 'snails', 'moist']

14
['published', 'book', 'novel', 'first', 'journal', 'written', 'series', 'newspaper', 'story', 'american', 'author', 'new', 'magazine', 'fiction', 'books', 'peerreviewed', 'also', 'science', 'publication', 'life']

13
['film', 'directed', 'starring', 'american', 'stars', 'released', 'written', 'based', 'drama', 'comedy', 'produced', 'also', 'films', 'silent', 'first', 'movie', 'roles', 'name', 'novel', 'documentary']

12
['album', 'released', 'band', 'records', 'first', 'studio', 'american', 'songs', 'music', 'second', 'release', 'recorded', 'rock', 'debut', 'live', 'tracks', 'label', 'albums', 'new', 'ep']

6
['navy', 'built', 'war', 'ship', 'uss', 'united', 'aircraft', 'class', 'world', 'states', 'launched', 'service', 'named', 'designed', 'first', 'royal', 'commissioned', 'american', 'ii', 'us']

4
['born', 'football', 'played', 'league', 'player', 'professional', 'plays', 'footballer', 'former', 'national', 'american', 'currently', 'also', 'hockey', 'rugby', 'team', 'australian', 'november', 'world', 'new']

2
['school', 'high', 'located', 'university', 'college', 'public', 'schools', 'students', 'education', 'county', 'district', 'founded', 'new', 'one', 'part', 'city', 'united', 'independent', 'established', 'catholic']

7
['historic', 'house', 'built', 'located', 'church', 'building', 'national', 'register', 'places', 'listed', 'county', 'street', 'united', 'known', 'also', 'museum', 'states', 'designed', 'added', 'hospital']

8
['river', 'lake', 'mountain', 'located', 'south', 'km', 'north', 'county', 'near', 'tributary', 'range', 'west', 'lies', 'creek', 'east', 'crater', 'state', 'ft', 'flows', 'pass']

11
['species', 'family', 'plant', 'genus', 'native', 'endemic', 'flowering', 'known', 'found', 'common', 'leaves', 'plants', 'habitat', 'tree', 'name', 'grows', 'orchid', 'south', 'bulbophyllum', 'perennial']

3
['born', 'american', 'known', 'new', 'band', 'writer', 'best', 'rock', 'music', 'work', 'musician', 'singer', 'also', 'york', 'author', 'books', 'album', 'united', 'university', 'guitarist']

5
['born', 'member', 'district', 'politician', 'state', 'house', 'democratic', 'senate', 'party', 'served', 'former', 'county', 'since', 'representatives', 'republican', 'united', 'elected', 'american', 'representing', 'national']

9
['village', 'district', 'population', 'province', 'located', 'census', 'municipality', 'nepal', 'india', 'state', 'county', 'km', 'people', 'within', '1991', '2010', 'township', 'central', 'southern', 'zone']

We observe that the results for ML (max likelihood) and priors with uniform distribution are exactly the same. This was expected as uniform distribution of priors provide no additional information and can be treated as a constant factor in calculation of posterior probabilities.

The overall accuracy improved slightly with ML classifier. We printed out the relative frequency (percentage of documents belonging to a class) of each class in training and testing data.
1
train:1.0 test:8.5
2
train:2.3 test:9.5
14
train:14.4 test:7.2
11
train:12.4 test:9.3
7
train:7.3 test:6.2
4
train:3.5 test:4.8
8
train:6.2 test:7.0
13
train:12.8 test:7.9
6
train:6.5 test:9.9
5
train:3.2 test:4.6
12
train:14.6 test:8.7
10
train:11.7 test:10.4
9
train:1.9 test:1.7
3
train:2.2 test:4.3

Hence, we observe that the test set is perhaps more uniformly distributed than the training set. At the very least, we do not see a tendency for test set data to obey the same distribution as the training set. Hence, the inclusion of prior probabilities slightly skews the prediction in a way that is not reflective of the test data.

The inclusion of class prior is not beneficial when the distribution of testing data is not known to be similar to the distribution of training data. The inclusion of priors would depend on the knowledge of distribution trends.

**Extra Credit:**
We implemented a bigram-based naive bayes text classifier, and a mixing procedure using the optional lambda_mix parameter in the provided TextClassifier.predict method.

We found that the performance of the bigram classifier was slightly worse that the unigram classifier. Bigram accuracy was 76.4%. We believe this is a logical finding, because given limited data, bigram classifiers miss important individual word features. For example, the word "fish" may occur often in documents about pets, but it may be preceded by a different word each time. A unigram model would correctly estimate that seeing "fish" in a test document increased the likelihood that that document is about pets. A bigram model, however, may see fish in a bigram that occured 1 or 0 times in the training data, and assign to it no special meaning. More generally, when we consider pairs of words and do not have a large amount of data, each pair of words will only occur rarely, providing less useful data for the classifier.

We predicted that scaling lambda from 0 to 1 would essentially scale performance from unigram to bigram performance. The was confirmed by testing: since bigram values were overall lesser in absolute value, applying a linear mix did not change any decisions for lambda values between 0 and 0.5. At 0.6, a small number of decisions were affected, and (by chance) performance increased very slightly. So (technically) 0.6 was the best lambda value, but it was only very slightly different from 0.0. After 0.6, decisions tended toward the decisions made by the bigram model, and at lambda = 1, performance was of course 76.4% as mentioned above.

**Statement of Contribution:**
Part 1 of the report was primarily written by Henry, and part 2 of the report was primarily written by Ayush. The extra credit section was written by Henry. Ayush and Henry both wrote code for all parts of the MP (except the extra credit which was only included in Henry's code) and compared their results. We submitted Henry's code for the autograder because it seemed to be slightly faster and included behavior for the lambda_mix extra credit optional parameter.