Dylan Parker

2. (c)

Optimal number of epochs (k=3) for the binary dataset: 19
Binary dataset training mistakes (k=3, epoch=19): 3
Binary dataset test mistakes (k=3, epoch=19): 18

Optimal number of epochs (k=3) for the count dataset: 48
Count dataset training mistakes (k=3, epoch=40): 46
Count dataset test mistakes (k=3, epoch=): 33

(d)

Binary dataset:
Most positive words: ['remov', 'sight', 'click', 'guarante', 'market', 'am', 'here', 'request', 'these', 'our', 'profession', 'pai', 'assist', 'most', 'monei']
Most negative words: ['wrote', 'releas', 'while', 'review', 'comment', 'someth', 'date', 'newslett', 'about', 'inc', 'execut', 'until', 'music', 'll', 'the']

Count dataset:
Most positive words: ['numberc', 'remov', 'anumb', 'report', 'face', 'dollarnumb', 'numberb', 'order', 'name', 'year', 'tv', 'market', 'guarante', 'will', 'numbera']
Most negative words: ['numberp', 're', 'if', 'wrote', 's', 'spam', 'but', 'file', 'url', 'date', 'server', 'd', 'about', 'cnet', 'last']

3. (a)

The Bernoulli Naive Bayes for the binary dataset had 58 mistakes. This is worse than using Perceptron for the binary dataset. The Multinomial Naive Bayes for the count dataset had 33 mistakes. This is about the same as using Perceptron for the count dataset. This indicates that Perceptron is better to use than Naive Bayes for spam email detection.

(b)

The logistic regression model for the binary dataset had 18 mistakes. This is about the same as using Perceptron for the binary dataset. The logistic regression model for the count dataset only had 23 mistakes. This is much better than using Perceptron for the count dataset. This indicates that using logistic regression performs much better than the other models and should be used when applicable.