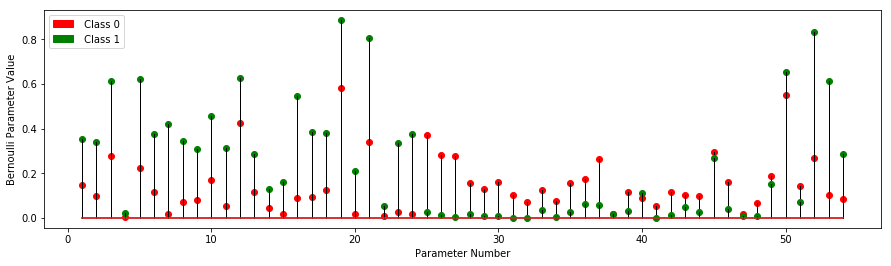
**2 (a): Confusion Matrix**

|  |  |  |
| --- | --- | --- |
| Class | Predicted 0 | Predicted 1 |
| Actual 0 | 54 | 2 |
| Actual 1 | 5 | 32 |

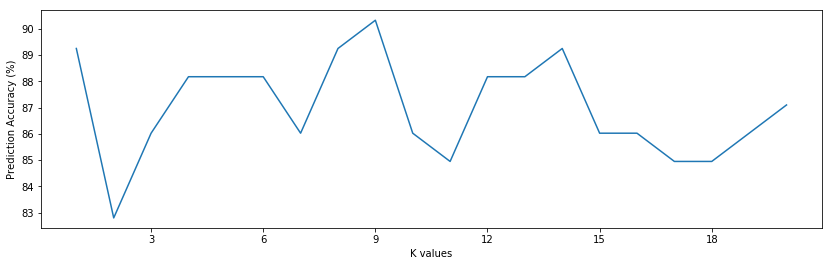
Prediction Accuracy: 92.473

**2 (b): Stem Plot**



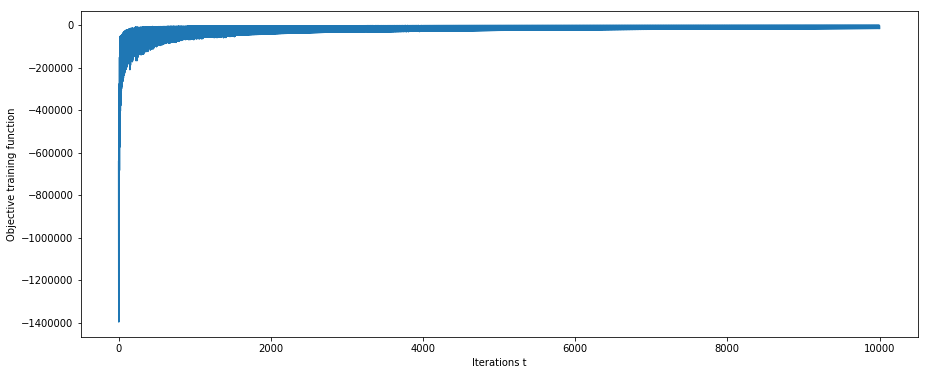
Looking at features 16 and 52, they represent the word ‘free’ and ‘!’ respectively. In both of these features the Bernoulli parameter for class 1 (Spam) represented by green dots on the stem plot are considerably larger than the Bernoulli parameter for class 0 (Not Spam). This indicates the word ‘free’ and the character ‘!’ occur much more frequently in spam emails than non-span emails. For instance the plot shows that the probability of seeing the word ‘free’ in spam emails is .545 compared to the probability of seeing it in non-spam emails being 0.0911. Similarly the probability of seeing the character ‘!’ in spam emails is 0.833 whereas the probability is only .269.

**2 (c): KNN Plot (Prediction Accuracy vs K)**

****

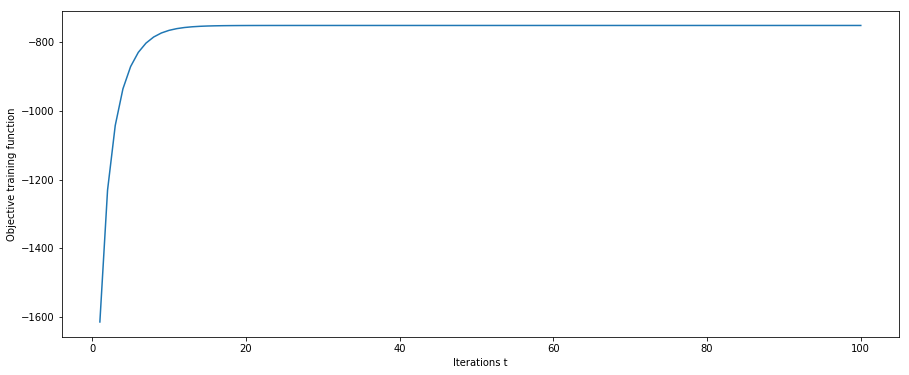
Ties were broken randomly in the KNN algorithm

**2 (d): Logistic Regression Plot**

****

Test Data Accuracy: 74.19%

**2 (e): Newton Method Plot**

****

Test Data Accuracy: 91.40%