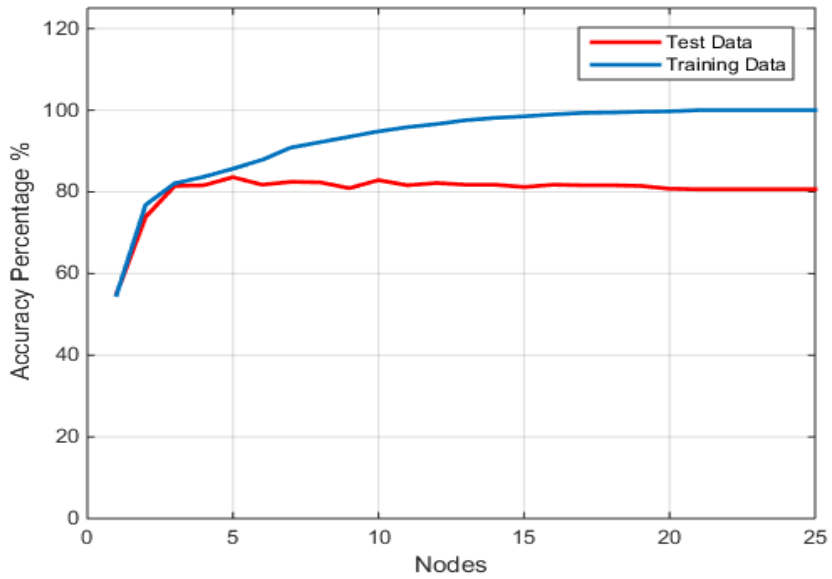
Assignment 3 Solutions

CS486/686 – Spring 2015

1.

- (20 marks)



- (10 marks) The over fitting happens after the fifth node that the accuracy of the decision tree for training data starts to decrease.
- (10 marks) The tree that achieved the highest testing accuracy is provided on the following page. For decision nodes the number after the word is the rounded information gain. “absent” means the word is not in the article and “present” means the word is in the article. At the leaf nodes ,”class 1” means the article is classified as atheism and “class 2” means the article is classified as graphics.
- (10 marks) For most of the words that are selected for the tree, it is possible to figure out intuitively why they are used. It is expected that some words like “graphics” or “image” would be associated with graphics, while words like “god” and “bible” should be associated with atheism. A rather unexpected set of words also exists in the tree, as well. For instance, the word “writes” is selected as the root node of the tree and is not a word that could be associated with either topic with high confidence. However, when these words are taken into account in combination with other words, the probability of having a certain type of document would increase significantly.

Tree with highest testing accuracy:

```
485 (writes) 0.2
    absent: 212 (god) 0.11
        absent: 153 (that) 0.08
            absent: 74 (bible) 0.05
                absent: class2 (graphics)
                present: class1 (atheism)
            present: 188 (wrote) 0.01
                absent: class2 (graphics)
                present: class1 (atheism)
        present: 184 (use) 0.21
            absent: class1 (atheism)
            present: 1 (archive) 1.0
                absent: class2 (graphics)
                present: class1 (atheism)
    present: 3143 (graphics) 0.12
        absent: 2109 (image) 0.086
            absent: 153 (that) 0.078
                absent: class1 (atheism)
                present: class1 (atheism)
            present: class2 (graphics)
        present: class2 (graphics)
```

2. a) (10 marks) **10 most discriminative words:**

(5 marks for words, 5 marks for opinion)

- graphics
- atheism
- religion
- moral
- evidence
- keith
- atheists
- god
- bible
- christian or religious (tied)

b) (10 marks) **Training and testing accuracies**

(5 marks each, full marks for within 1%, part marks for within 3%)

Training accuracy: 92.84%

Testing accuracy: 88.97%

c) (10 marks) **The naive Bayes model assumes that all word features are independent. Is this a reasonable assumption? Explain briefly.**

It is possible to argue for or against the assumption. Below are two possible answers:

Unreasonable assumption: Words are not actually independent. For example, a document containing the word 'bible' may be more likely to contain the word 'religion'.

Reasonable assumption: The independence assumption, though not entirely accurate, works well enough in practice. In our results we see that the testing accuracy of the naïve Bayes model under this assumption is better than the decision tree learner.

d) (10 marks) **What could you do to extend the Naive Bayes model to take into account dependencies between words?**

One possible extension is to add arcs between word features. Other reasonable ideas were also accepted.

e) (10 marks) **Which approach performs best among decision trees and the naive Bayes model? Explain briefly why.**

The naïve Bayes model performed better than the decision tree. Part of the reason may be that the decision tree begins to overfit quickly. We can see this by some of the chosen words in the decision tree such as 'that' and 'use'.

Note: If your results showed that the decision tree performed better you were not penalized in this question as long as your explanation was reasonable.