CSE 142: HW3 Writeup
Naïve Bayes and Decision Trees

Amir Noori

05/08/25

1   Introduction

In this assignment, I implemented two classifiers from scratch: Naïve Bayes and Decision Trees. I applied them to two datasets: Mushroom (to predict whether a mushroom is edible) and Congressional Voting Records (to predict political party). The goal was to understand how these models work internally and compare their performance to the scikit-learn versions. The task included training, testing, tuning, and evaluating these models.

2   Naïve Bayes Classifier

The Naïve Bayes model was built from scratch using basic probability with Laplace smoothing. I wrote three functions:

- `prior_prob`: estimates the class priors $P(y)$
- `like`: estimates the likelihoods $P(x_j \mid y)$ for each feature
- `NB`: makes predictions using the log of the product of probabilities

Results

Dataset	Accuracy (%)
Mushroom	95.84
Voting	91.61

Confusion Matrices

Mushroom

	Pred 0	Pred 1
Actual 0	395	6
Actual 1	19	362

Voting

	Pred 0	Pred 1
Actual 0	41	9
Actual 1	4	39

3 Decision Tree Classifier

I built a recursive tree-building function based on entropy and information gain. If all labels were the same or the max depth was reached, a leaf was returned.

- `entropy`: computes the class impurity
- `info_gain`: calculates the gain for a split
- `induce_tree`: builds the tree recursively
- `get_tree`: predicts using the built tree

The model was evaluated at multiple depths using the dev set to find the best one.

Results

Dataset	Accuracy (%)
Mushroom (best depth = 4)	100.00
Voting (depth = 4)	95.16

Confusion Matrices

Mushroom

	Pred 0	Pred 1
Actual 0	401	0
Actual 1	0	381

Voting

	Pred 0	Pred 1
Actual 0	46	4
Actual 1	4	39

4 Sklearn Models

For comparison, I also used `CategoricalNB` and `DecisionTreeClassifier` from scikit-learn. I set the depth of the decision tree to the same value as my best model (4).

Results

Model	Accuracy (%)
NB (Mushroom)	95.84
DT (Mushroom)	100.00
NB (Voting)	91.61
DT (Voting)	95.16

5 Conclusion

Both the Naïve Bayes and Decision Tree implementations were successful. Decision Trees performed slightly better, especially for the Mushroom dataset, where it reached perfect accuracy. Naïve Bayes was more efficient but had slightly lower accuracy. Comparing with scikit-learn, my results were very close, which shows that the implementations were correct. The dev set was useful for tuning tree depth. Overall, this assignment helped solidify my understanding of both classifiers and gave me experience building them from scratch and analyzing real datasets.