Cory Parker

IDAI610-Naïve Bayes

1. Laplace smoothing adjusts the maximum likelihood estimation to account for the presence of features in the test data that may not have been observed in the training data. By adding 1 to the count of each feature and adjusting the denominator accordingly, we prevent the calculated probabilities from being zero, which can skew the results. This makes the model more robust and better at handling unseen data.
2. Preprocessing: I converted all the text to lowercase, eliminated punctuation and removed stop words using the *nltk* library and removed HTML tags, for tokenization of text by splitting the document into a list of words, returning only the words not in the stop words.
3. To compare results on tests for both datasets, I created a confusion matrix, presenting counts of the true positive, false positive, true negative, and false negative predictions across classes, with actual labels on the y-axis and predicted labels on the x-axis.
4. When analyzing differences, the performance between datasets, I considered class distribution, feature representation, and the nature of the text. If there’s an imbalance of classes, this could lead to differences in performance as well. Visualizations help with this as well as considering context.
5. Naïve Bayes algorithm is computationally efficient and performs well with high-dimensional data and is simple to implement. The assumption of independence might be unrealistic in the real world, but we get surprisingly accurate results, none the less.
6. Laplace Smoothing is a technique used to handle the issue of zero probability in Naïve Bayes classifiers. It’s also useful for handling out of vocabulary words…
7. Conditional Independence in Bayes Networks refers to the relationship between variables where the state of one variable is independent of the state of the another given the state of a third variable. Wet\_Grass is independent of Cloudy, given both states of Sprinkle and Rain, since it cannot be determined by whether it is raining or the sprinkler is on, regardless of whether it is cloudy. Sprinkler may be conditionally independent of Rain given Cloudy. The conditional independencies are represented in the structure of the graph and the conditional probability tables (CPTs) associated with each node.