Bayes Net Structure Prediction of a Text dataset

• Application Domain

The application domain is primarily increasing the accuracy of a text/Document classifer by learning the Bayes Net Structure particularly for Text Dataset .

• Problem to tackle

In most cases the Naive Bayes Model is followed which can lead to less accurate models in many cases. Knowing the dependencies between different features in te dataset can help build better models for Text/Document Classification. As per the Naive Bayes Model for Text Data , all the features are assumed to be conditionally independent given the label. This works well enough for Spam Classification but for cases where the labels are dependent of the meaning of the text, the naive Bayes Model does not perform well. Hence, there is a need to know how the features are dependent on each other.

• Artificial Intelligence techniques to use

This project will primarily involve the usage of Machine Learning and Natural Language Processing techniqes along with Text Preprocessing such as Lemmatization, Stemming along with other text preprocessing techniques.

References

[1] C. Chow and C. Liu. Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, 14(3):462–467, May 1968.

[2] Nir Friedman, Dan Geiger, and Moises Goldszmidt. Bayesian network classifiers. *Machine learning*, 29(2-3):131–163, 1997.

[3] Liangxiao Jiang, Zhihua Cai, Dianhong Wang, and Harry Zhang. Improving tree augmented naive bayes for class probability estimation. *Knowledge-Based Systems*, 26:239–245, 2012.

[4] Liangxiao Jiang, Harry Zhang, Zhihua Cai, and Jiang Su. Learning tree augmented naive bayes for ranking. In *Database Systems for Advanced Applications*, pages 688–698. Springer, 2005.

[5] Ron Kohavi. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *KDD*, volume 96, pages 202–207. Citeseer, 1996.

[6] Fei Zheng and Geoffrey I. Webb. *Encyclopedia of Machine Learning*, chapter Tree Augmented Naive Bayes, pages 990–991. Springer US, Boston, MA, 2010.