ML 1 - Yilmaz

2022-23

11/28/2022

**Quarter 2 Project Proposal**

**Group Members:** Isabella Zhu, Lilian Zhu

**Motivation / Problem Statement**

Naive Bayes is a probabilistic classifier algorithm that achieves fairly high accuracy levels. However, Naive Bayes makes the assumption that all attributes are independent of each other, which is not true of most real life data. When this assumption doesn’t hold, Naive Bayes decreases in accuracy. Our goal is to build upon the Naive Bayes algorithm, increasing the accuracy when the attributes are correlated without sacrificing accuracy when the attributes aren’t correlated.

**Method**

We propose an improvement on the Naive Bayes algorithm that takes into account correlation between two attributes.

Naive Bayes assumes that the probability of two events occurring is as follows:

This is only true if and are independent events, which is the “naive” assumption that Naive Bayes makes. We generalize the formula so it applies to any and , regardless of whether they are independent or not.

where is the correlation coefficient between and .

To calculate for instance and class , we will take the product of the pairwise given by the generalized formula.

**Intended Experiments**

We will compare our algorithm against Naive Bayes, Hidden Naive Bayes, and Generalized Naives Bayes (which utilizes a weighting function). We will test the algorithms on two datasets: COVID-19 Dataset and credit-g. COVID-19 Dataset has independent attributes and credit-g has dependent attributes. Our goal is to have a higher accuracy than Naive Bayes on credit-g without having a lower accuracy on COVID-19. Note that both datasets are binary (two class labels) because GNB only works on binary datasets.

We will evaluate the performance of our algorithm by calculating accuracy and confusion matrices for each dataset. We will also calculate these metrics with Naive Bayes, Hidden Naive Bayes, and Generalized Naive Bayes to compare results. Finally, we will evaluate the efficiency of each algorithm by analyzing big-O and comparing runtimes for both creating the model and testing the model.

**Datasets**

The dataset credit-g was included in the initial WEKA download.

The COVID-19 dataset was created by the Mexican government and posted [on their website](https://datos.gob.mx/busca/dataset/informacion-referente-a-casos-covid-19-en-mexico) by Hector Parades. A translation of this dataset can be found posted on [Kaggle](https://www.kaggle.com/datasets/meirnizri/covid19-dataset) by user Meir Nizri.

**Prior Research**

Hidden Naive Bayes is an algorithm that improves on Naive Bayes by creating “hidden” parent nodes that encapsulate dependent probabilities, so it bypasses Naive Bayes’ assumption of independence between attributes. Each attribute has a hidden parent attribute. Weights are calculated for each pairwise combination of attributes. Then, the weighted influence of each attribute is encapsulated in each parent.

Source: [Hidden Naive Bayes](https://www.aaai.org/Papers/AAAI/2005/AAAI05-145.pdf)

Generalized Naive Bayes is an algorithm that improves on Naive Bayes by relaxing the independence assumption. It does this by creating a smooth function for each attribute which accounts for marginal bias from other attributes. GNB also uses kernels to deal with ties when predicting instances. These kernels are produced given a smoothing factor, ƛ. However, GNB only works with binary datasets (i.e. two class labels).

Source: [Naives Bayes with Weight Function](https://dl.acm.org/doi/pdf/10.1145/1089815.1089826)

**Sources / Links to related papers**

* <https://link.springer.com/content/pdf/10.1007/s10994-005-4258-6.pdf?pdf=button>
* [Naives Bayes with Weight Function](https://dl.acm.org/doi/pdf/10.1145/1089815.1089826)
* [Hidden Naive Bayes](https://www.aaai.org/Papers/AAAI/2005/AAAI05-145.pdf)
* <https://math.stackexchange.com/questions/1205928/probability-of-three-events-occurring-given-correlation>