Enhancing Naive Bayes Algorithm with Stable Distributions for Classification

Supplementary Material

Nahush Bhamre, Pranjal Prasanna Ekhande, and Eugene Pinsky

Department of Computer Science, Metropolitan College, Boston University,
1010 Commonwealth Avenue, Boston, MA 02215
nahush@bu.edu, pekhande@bu.edu, epinsky@bu.edu (corresponding author)

6 Brief Description of Datasets

- **Banknote Authentication** [1]: Used for authenticating banknotes based on features extracted from images of the currency, including wavelet-transformed features for classification tasks.
- **Blood Transfusion** [2]: Contains data related to blood donation, used for predicting whether a blood donor will donate within a given time window based on historical donation patterns.
- **Breast Cancer**[3]: Used for predicting breast cancer recurrence based on attributes from patient biopsies.
- **Customer Churn [4]**: Focuses on predicting customer churn using telecommunications data, assessing factors such as customer usage patterns and service changes. 8 out of 13 features were selected which were continuous variables.
- **Diabetes** [5]: Commonly used to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements.
- **Electrical Grid Stability** [6]: The local stability analysis of the 4-node star system (electricity producer is in the center) implementing Decentral Smart Grid Control concept.
- **Heart Disease** [7]: Used to predict the presence of heart disease in patients based on various medical attributes, including age, sex, blood pressure, and cholesterol levels. 5 out of 13 features were selected which were continuous variables.
- **Image Segmentation**[8]: Contains instances drawn randomly from a database of 7 outdoor images. The images were hand-segmented to create a classification for every pixel.
- **Occupancy Estimation [9]**: This dataset includes sensor data from an office room and is used for classifying occupancy status based on environmental conditions like temperature and humidity.
- **Rice Dataset** [10]: Used for classifying different types of rice kernels based on features measured from images of the kernels.
- **Seeds Dataset** [11]: Provides measurements of wheat seed varieties, often used for classification of different types of wheat kernels based on their geometric properties.
- **Smoke Detection (IoT)** [12]: Contains air quality sensor data that helps in detecting the presence of smoke, suitable for IoT-based smoke detection applications.
- **Sonar** [13]: The Sonar dataset is used for binary classification, specifically to distinguish between sonar signals reflected by metal and those by rocks.
- **Statlog (Vehicle Silhouettes)** [14]: The purpose is to classify a given silhouette of different types of vehicle, using a set of features extracted from the silhouette.

– **Water Potability** [15]: Assesses water quality based on chemical properties, allowing for classification of water samples as potable or not.

References

1. V. Lohweg. ”Banknote Authentication,” UCI Machine Learning Repository, 2012. [Online]. Available: https://doi.org/10.24432/C55P57.
2. I. Yeh. ”Blood Transfusion Service Center,” UCI Machine Learning Repository, 2008. [Online]. Available: https://doi.org/10.24432/C5GS39.
3. W. Wolberg, O. Mangasarian, N. Street, and W. Street. ”Breast Cancer Wisconsin (Diagnostic),” UCI Machine Learning Repository, 1993. [Online]. Available: https://doi.org/10.24432/C5DW2B.
4. ”Iranian Churn,” UCI Machine Learning Repository, 2020. [Online]. Available: https://doi.org/10.24432/C5JW3Z.
5. M. Kahn. ”Diabetes,” UCI Machine Learning Repository, [Online]. Available: https://doi.org/10.24432/C5T59G.
6. V. Arzamasov. ”Electrical Grid Stability Simulated Data ,” UCI Machine Learning Repository, 2018. [Online]. Available: https://doi.org/10.24432/C5PG66.
7. A. Janosi, W. Steinbrunn, M. Pfisterer, and R. Detrano. ”Heart Disease,” UCI Machine Learning Repository, 1989. [Online]. Available: https://doi.org/10.24432/C52P4X.
8. ”Image Segmentation,” UCI Machine Learning Repository, 1990. [Online]. Available: https://doi.org/10.24432/C5GP4N.
9. A. Singh and S. Chaudhari. ”Room Occupancy Estimation,” UCI Machine Learning Repository, 2018. [Online]. Available: https://doi.org/10.24432/C5P605.
10. ”Rice (Cammeo and Osmancik),” UCI Machine Learning Repository, 2019. [Online]. Available: https://doi.org/10.24432/C5MW4Z.
11. M. Charytanowicz, J. Niewczas, P. Kulczycki, P. Kowalski, and S. Lukasik. ”Seeds,” UCI Machine Learning Repository, 2010. [Online]. Available: https://doi.org/10.24432/C5H30K.
12. Blattmann, S. (2023). Smoke Detection Dataset [Online]. Kaggle Machine Learning Repository. https://www.kaggle.com/datasets/deepcontractor/smoke-detection-dataset.
13. T. Sejnowski and R. Gorman. ”Connectionist Bench (Sonar, Mines vs. Rocks),” UCI Machine Learning Repository, 1988. [Online]. Available: https://doi.org/10.24432/C5T01Q.
14. P. Mowforth and B. Shepherd. ”Statlog (Vehicle Silhouettes),” UCI Machine Learning Repository, [Online]. Available: https://doi.org/10.24432/C5HG6N.
15. ”Water Potability” [Dataset]. Kaggle Machine Learning Repository. [Online]. Available: https://www.kaggle.com/datasets/adityakadiwal/water-potability