

Naïve bayes-2

Assignment Questions



Q1. A company conducted a survey of its employees and found that 70% of the employees use the company's health insurance plan, while 40% of the employees who use the plan are smokers. What is the probability that an employee is a smoker given that he/she uses the health insurance plan?

Q2. What is the difference between Bernoulli Naive Bayes and Multinomial Naive Bayes?

Q3. How does Bernoulli Naive Bayes handle missing values?

Q4. Can Gaussian Naive Bayes be used for multi-class classification?

Q5. Assignment:

Data preparation:

Download the "Spambase Data Set" from the UCI Machine Learning Repository (<https://archive.ics.uci.edu/ml/datasets/Spambase>). This dataset contains email messages, where the goal is to predict whether a message is spam or not based on several input features.

Implementation:

Implement Bernoulli Naive Bayes, Multinomial Naive Bayes, and Gaussian Naive Bayes classifiers using the scikit-learn library in Python. Use 10-fold cross-validation to evaluate the performance of each classifier on the dataset. You should use the default hyperparameters for each classifier.

Results:

Report the following performance metrics for each classifier:

Accuracy
Precision
Recall
F1 score

Discussion:

Discuss the results you obtained. Which variant of Naive Bayes performed the best? Why do you think that is the case? Are there any limitations of Naive Bayes that you observed?

Conclusion:

Summarise your findings and provide some suggestions for future work.

Note: This dataset contains a binary classification problem with multiple features. The dataset is relatively small, but it can be used to demonstrate the performance of the different variants of Naive Bayes on a real-world problem.

Note: Create your assignment in Jupyter notebook and upload it to GitHub & share that github repository link through your dashboard. Make sure the repository is public.